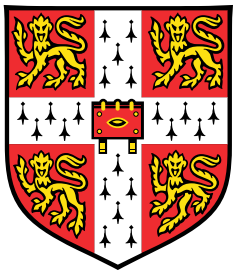
Detection of trace levels of circulating tumour DNA in early stage non-small cell lung cancer



Katrin Heider

Supervisor: Dr Nitzan Rosenfeld

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

King's College September 2019

Dedicated to my grandfather. Thank you for inspiring me to ask questions and understand the world around me. I miss you.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Katrin Heider
September 2019

Acknowledgements

My PhD has been an adventurous four year journey in which I learned a lot, not just about science but also about myself. It has been an absolute pleasure to work alongside some of the smartest people I have ever met and I am grateful for all the help and feedback I have received in the process. I remember first coming into the lab and feeling entirely underqualified for what I was supposed to be doing there. Thanks to the patience of all the Rosenfeld lab members I was able to develop the necessary skills to become a ctDNA expert, and I even picked up bioinformatics on the side.

First and foremost I would like to thank Nitzan Rosenfeld for sharing his wisdom, supporting me in my endeavours, and providing me with the space and funds needed for the projects. Nitzan's background and expertise in both, the academic and industrial setting, have provided me with a unique insight to how people from different areas can look at the same biological problem from quite different angles. Thank you Nitzan for sharing your expertise with me and giving me the trust to complete my PhD rather independently in your lab.

I would also like to thank the rest of the Rosenfeld lab. Charlie Massie, who was my supervisor during my first year, has helped me understand the requirements for a translational project and all the paperwork involved. He has taught me in and outside the lab, thereby laying the foundation for the following years. Chris Smith and Florent Mouliere have both been an amazing resource when it came to general questions around the lab or the project. Their help and support when I set out to write my first (and potentially only) book chapter has been much appreciated and it would not have been possible without them. Wendy Cooper, Irena Hudecova, Angela Santonja Climent and Davina Gale are all not only amazing examples of women in science but have also been very supportive whenever I was struggling, either in the lab or with one of the many collaborators. James Morris and Dineika Chandrananda have taken the time to teach me bioinformatics from scratch. Thanks to those two I now know how to open the terminal and what I can do with it. Of course their support went way further than that and both have helped me greatly throughout the years. Finally, I would like to thank all the other members of the Rosenfeld lab for their help along the way and for making the lunch breaks so enjoyable.

A very special thank you goes to Jonathan Wan. When I first learned we were going to collaborate on a daily basis I was not too keen, mostly because the two of us function quite differently and I could not imagine how that was going to work out in a close collaboration. However, working on INVAR with you has been one of the most fun experiences in my PhD and I am beyond amazed at what we have accomplished together. Your willingness to go the extra mile combined with your expertise have made collaborating very exciting and easy. Along the process we not only learned how to communicate effectively and share the work logically but you have also become a close friend that I truly enjoyed working with.

Through the LUCID study I met the clinicians and associated staff, which provided me with yet another view on science. Thank you Robert Rintoul, Susan Harden and Doris Rassl for taking the time and explaining the clinical importance of our work to me. Our discussions have been enjoyable and very useful. I would also like to thank all the members of the Cambridge Clinical Trial Center for their support. Without your help the LUCID study would not have been feasible and we would have been drowning in paperwork.

Outside the lab I have had an amazing group of friends who helped to keep me grounded. I would like to thank the German Stammtisch for bringing a bit of German culture into my life in England and distracting me from the lab work when it was needed. I would like to thank Cora for always having an open door and supporting me in my various endeavours here in Cambridge. Together with Caro and Anne we made some good memories during our time here. Thank you also to Cello for our many discussions and for teaching me a thing or two about aquariums and 3D printers. I would also like to thank the Parkside Pool. The pool has and probably always will be a place for me to detach. To properly disconnect from everything and just be with myself. Continuing to swim during my PhD turned out to be a great choice. Not only did I find the love for swimming again, it also brought me some of my closest friends during my time here. Thank you James for giving me the opportunity to join your team and thank you Kathryn for not only becoming a great friend but also pushing me to come to training when I needed some pushing.

And finally, thank you to my parents Wolfgang and Ute, my brother Felix and the rest of my family for supporting me from main-land Europe. You have been the rock in my life from day one. I would have never made it this far without you. Thank you for your endless love, support and understanding.

Abstract

Liquid biopsies, using analytes such as circulating tumour DNA (ctDNA), can detect and quantify cancer in a minimally invasive way and can provide information on tumour heterogeneity. Current limitations in the liquid biopsy field are centred around the general sensitivity of the present assays and the input and logistical requirements for sensitive detection.

While detection rates from liquid biopsy platforms are good when ctDNA levels and tumour burden are high, they are lacking the required sensitivity to detect cancer in the early stage and minimal residual disease setting. Here I present the INtegration of VAriant Reads (INVAR) pipeline, which can greatly enhance the sensitivity of ctDNA detection by utilising large patient specific mutation lists on sequencing data. The methodology can be applied to (custom) capture and (shallow) whole genome sequencing data and detects ctDNA down to parts per million, proving more sensitive than previously published methods. INVAR was applied to samples from 90 treatment naïve patients with early stage non-small cell lung cancer to characterise the ctDNA levels and could provide a better sensitivity than comparable cohorts. In addition, I outsourced samples from 27 patients to be analysed with the InVisionSeq™ assay, which does not require a priori tumour information but proved to be less sensitive than INVAR.

I also assessed the potential for reducing the logistical burden in ctDNA analysis by sampling minimal blood volumes that do not require immediate processing. I interrogated if ctDNA can be detected from as little as a blood spot and show ctDNA detection using this approach in both, xenograft and human samples. ctDNA detection from blood spots provides a means to sample xenograft models without having to sacrifice the animal, allowing for longitudinal monitoring in this setting. It may also present an opportunity to frequently sample blood from patients and could reduce the logistical burden on sample collection and processing. In the future, after optimising the protocol, this could serve to reduce the complexity of clinical/translational studies. The removal of large DNA fragments using this protocol may also facilitate the analysis of ctDNA from archival cohorts where samples were collected under suboptimal conditions.

Lay Abstract

Cells can release DNA into the blood, either passively when they die or actively when alive. In cancer patients, the cancerous cells also release DNA into the circulation, called circulating tumour DNA (ctDNA). By obtaining a blood sample from a cancer patient, researchers can obtain information on the cancer from this ctDNA, without the need to analyse a surgical specimen or a biopsy collected from the tumour by a large needle. While this minimally invasive method is very promising for the analysis of cancer, the sensitivity of currently used analysis methods is limited especially for the detection of early cancer or of low disease burden potentially remaining after initial treatment. Additionally, the current best practices for plasma collection require prompt processing or the use of cell-preserving materials and restrict the broad application and study of ctDNA.

The level of ctDNA in the circulation correlate with the tumour size of the patient. Patients with late stage disease and larger tumour volumes tend to have a higher concentration of ctDNA, making its detection easier, while patients with early stage disease or patients after treatment have smaller tumour volumes and lower levels of ctDNA, making its detection more difficult. The currently popular methods in the field target a fixed number of genes known to be related to cancer or few cancer specific mutations and have been effective in detecting high levels of ctDNA. However, they lack sensitivity in detecting low ctDNA levels. Here, I aim to improve sensitivity by targeting more point mutations and developed a method called INtegration of VAriant Reads (INVAR). INVAR is generalisable to a variety of sequencing data and can detect ctDNA at levels up to 100 times lower than other methods. When applied to 90 patients with early stage lung cancer, INVAR could detect ctDNA in the majority of patients, showing the potential of the INVAR method.

In this work I also present a method for sample processing that requires minimal processing of samples at the time of collection. This can simplify protocols and allow more extensive collection of samples for analysis. In this case a drop of blood is applied to a filter paper and left to dry. DNA can then be extracted from the dried blood spot and analysed for the presence of ctDNA. In the first application ctDNA was detected in both, a human blood sample and blood from a mouse that had an implanted human tumour. Detection from a dried blood spot opens opportunities to collect samples at high frequency from animals with implanted

human tumours, which was previously difficult due to their lower blood volume. It also offers the possibility to collect patient samples more frequently and potentially even allows for sample collection at home. In the future, once the method has been optimised further, ctDNA analysis from dried blood spots could reduce the complexity of studies. Additionally, the same method could also be used to analyse older samples that were collected under suboptimal conditions and are not suitable for analysis with current methods, and may be 'rescued' for study using this new method.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AF	Allelic Fraction
AMP	Anchored Multiplex PCR
AVAST − M	Adjuvant bevacizumab in patients with melanoma at high risk of recurrence
bp	basepairs
CAPP − Seq	CAncer Personalized Profiling by Deep Sequencing
cfDNA	cell free DNA
CNA	Copy Number Alteration
CT	Computed Tomography
ctDNA	circulating tumour DNA
FFPE	Formalin-Fixed Paraffin-Embedded
GP	General Practice
H&E	Haemotoxylin and Eosin
hGA	Haploid genomes analysed
iDES	integrated Digital Error Suppression
INVAR	Integration of Variant Reads
IR	Informative Reads
LUCID	LUng cancer CIrculating tumour Dna Study

MAF	Mutant Allele Fraction
MDA	Multiple Displacement Amplification
MELR	MelResist
mFast − SeqS	modified Fast Aneuploidy Screening Test-Sequencing System
NSCLC	non-small cell lung cancer
PCR	Polymerase Chain Reaction
PDX	Patient Derived Xenograft
ROC	Receiver Operating Characteristic
Safe − SeqS	Safe-Sequencing System
SNV	Single Nucleotide Variant
SV	Structural Variant
sWGS	shallow Whole Genome Sequencing
t − MAD	trimmed median absolute deviation from copy number neutrality
TAm − Seq	Tagged-Amplicon Sequencing
TAPAS	TAilored-Panel Sequencing
TEC − Seq	Targeted Error Correction SEQuencing
TRACERx	TRAcking Cancer Evolution through therapy (Rx)
UMI	Unique Molecular Identifier
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Chapter 1

Introduction

1.1 Attribution

Section 1.3 in this chapter is adapted from the book chapter ”Overview of selected approaches for cell free DNA library preparation and sequencing”, which was submitted in January 2018 as part of the book “Circulating Tumour DNA – Purification and Analysis Techniques”:

“Overview of selected approaches for cell free DNA library preparation and sequencing”
K. Heider[§], F. Mouliere, C. G. Smith[§]
[§] Corresponding author

Part of section 1.2 in this chapter was previously included in my first year report.

1.1.1 Author contributions

The book chapter and its associated figures were prepared by me. I have received feedback from both Florent Mouliere and Chris Smith along the way. The first year report was written by me, with feedback from Charlie Massie. It was initially written and submitted in July 2016 and was part of the first year evaluation process.

1.2 The biology of cell free DNA

Circulating cell free DNA (cfDNA) was first described in 1948 by Mandel and Metais [1] and has been of interest in many different areas of research such as prenatal diagnostics [2], athletics [3], transplantation patients [4], and acute trauma patients [5]. cfDNA and its possible importance for cancer patients was first described in 1977 by Leon et al. [6].

cfDNA has a multi-modal size distribution and is found as short 70 – 200bp fragments but also as long 21kb fragments [7]. The difference in fragment length is believed to be due to the mechanism of its release by the cells of origin. The shorter fragments are likely to be released through apoptosis, while the long fragments are thought to be associated with exosomes or cells undergoing necrosis (see Fig. 1.1A) [3, 8]. The length of plasma cfDNA fragments centres at 166bp, representing the length of DNA wrapped around the nucleosome and a linker histone (see Fig. 1.1B) [9]. This mode of distribution is usually complemented by additional peaks at multiples of 166bp, representing di- and trinucleosomes, typical of an apoptotic ladder [10, 11]. Additionally, cfDNA shows a 10bp oscillatory pattern below fragment lengths of 150bp, likely related to the length of individual helical turns as the DNA winds around the histone (Fig. 1.1B) [9, 12].

In cancer research, scientists are interested in cfDNA that is released specifically by tumour cells (called circulating tumour DNA, or ctDNA) [13, 8]. Compared to cfDNA, ctDNA has been shown to be even more fragmented, exhibiting a mode of distribution between 133-145bp (or smaller) in length [14, 15]. The tumour derived ctDNA carries information on the genetic changes of the tumour and can guide treatment decisions [16].

ctDNA has been demonstrated to represent most of the cellular populations that make up a given tumour. As such it allows a global ‘snap-shot’ of the disease at a given time. Conversely, conventional biopsy techniques might not capture this global representation [17–19]. This is especially important for late stage disease where the assessment of metastasis via biopsy becomes technically challenging. Additionally, the ease of sampling and minimally invasive nature of cfDNA collection allow for serial monitoring of patients, which would otherwise be harmful using biopsies or computed tomography (CT) scans [20, 21]. In the long run, the aim is to broaden the clinical applications of ctDNA as a biomarker, helping to diagnose cancer patients earlier, monitor treatment response more effectively, and stratify patients according to risk groups or specific treatments [22, 23].

Our understanding of cfDNA in the cancer setting, and ways to analyse it, have improved greatly since it was first described [6, 24]. The first diagnostic test using plasma ctDNA to identify non-small cell lung cancer (NSCLC) patients with an EGFR exon19 deletion has already been FDA-approved for the clinic. This alteration is important as patients carrying this mutation can be effectively treated with the tyrosine kinase inhibitor gefitinib (Iressa) [16].

More recently, the COBAS® EGFR mutation test v2 was FDA approved for using plasma samples of NSCLC patients to detect alterations that allow treatment with the tyrosine kinase inhibitors erlotinib (Tarceva) and osimertinib (Tagrisso) [25]. Multiple studies have also shown the utility of ctDNA as a means of identifying recurrence sooner than conventional methods, with important clinical implications [19, 26]. Finally, with the development of easy-to-use kits and the continuous reduction in sequencing costs, the analysis of ctDNA in the clinic represents an increasingly viable and attractive option. Recent developments that tailor sequencing to the biology of ctDNA, innovative sequencing solutions, as well as integration of machine learning analysis of large scale cohorts will also lead to intriguing new avenues for implementation of liquid biopsy in the clinic.

1.2.1 Challenges associated with circulating tumour DNA analysis

The concentration of cfDNA in the blood varies between 5 and 10 ng/mL of plasma [28, 29]. However, only a small proportion of this will be ctDNA released by the tumour. Depending on the cancer stage and subtype, this proportion can vary quite substantially. Bettegowda and colleagues showed increased levels of ctDNA at later disease stages: patients with stage IV disease showed a median ctDNA concentration of around 10% while patients with early stage disease showed only a few copies of mutant ctDNA in the circulation [30]. Overall, levels of ctDNA in the circulation correlate with tumour burden; patients with a larger tumour size tend to also have higher levels of ctDNA [19, 31, 32]. The more cancerous cells present in the body, the more cancerous DNA can be released into the circulation. Interestingly, they also observed great variability with regards to ctDNA concentration within each stage, indicating that stage alone is not the only factor affecting ctDNA concentration in the circulation. The same study also analysed ctDNA detection rates in different cancer types and could show that the detectability of ctDNA with a PCR-based assay or Safe-SeqS, a massively parallel sequencing method, was close to 100% in the plasma of advanced bladder, colorectal, gastroesophageal, and ovarian cancer patients [30]. However, for advanced glioma or kidney cancer patients, the average rate of detection was only around 10%, highlighting the effect of cancer type on the ability to detect ctDNA [30]. Therefore, both disease stage and cancer type are affecting the ability of detecting ctDNA. Later studies could also identify differences in detection rates with respect to cancer subtypes [19].

The majority of ctDNA analysis assays have focused on identification of tumour specific single nucleotide variants (SNVs) but efforts have also been made by exploring methylation profiling [33] and structural rearrangement assays [34]. When focusing on SNVs, the field has been using both, PCR and sequencing based methods.

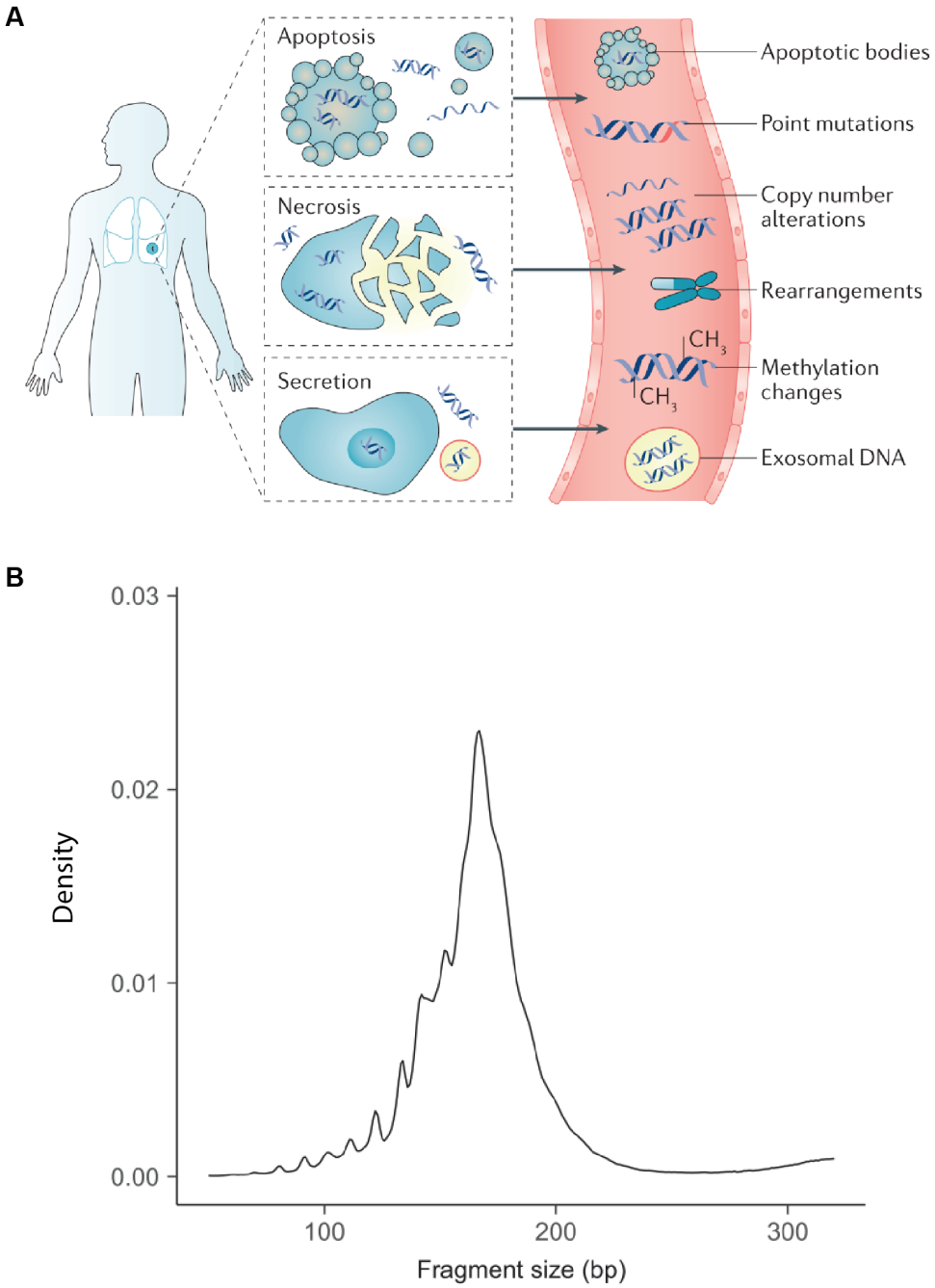


Fig. 1.1 **Cell free DNA characteristics.**
(**A**) Mechanisms of cfDNA release. cfDNA is thought to be released via a variety of mechanisms such as apoptosis, necrosis or active secretion. Taken from Wan et al. [8]. (**B**) Size profile of cfDNA. A characteristic peak around 166bp fragment length corresponds to the DNA wrapped around the histone and its linker. An oscillatory pattern is observed for fragments below 150bp and corresponds to individual DNA turns around the histone. Adapted from Mouliere et al. [27].

Once sequencing became cheaper, efforts have been made to increase the number of interrogated loci in a given sample, thereby increasing the overall sensitivity and chance of detection. Fig. 1.2 depicts this concept: If one were to target only a single mutant locus, the maximum sensitivity of a perfect assay is equivalent to the total number of cfDNA input genomes (G). The total number of input genomes (G) is limited by the volume of blood realistically obtainable from a cancer patient (usually between 2 and 10mL). As a result, the potential sensitivity from a single locus assay is limited and can never be better than a few molecules per G. If one were to interrogate multiple mutated loci (N) one can now interrogate the G input genomes across the N mutations (NxG). This improves the sensitivity of the assay and becomes a few molecules per NxG. This concept of targeting multiple mutations has proven especially useful when disease burden and ctDNA concentrations are low [8, 19, 31, 35, 36].

The main challenges in the ctDNA field is the low proportion of ctDNA in the circulation. Depending on the stage and tumour type of the patient, the detection of ctDNA can become quite difficult. As illustrated in Fig. 1.2, the detection of ctDNA is directly related to the total number of available molecules for the analysis. This in turn can be understood as the product of the total number of mutated loci targeted and the amount of genetic material available for the analysis. Therefore, one can optimise the chances to detect ctDNA in any given setting by trying to maximise one or both of these parameters.

1.3 Sequencing based methods for cfDNA analysis

There are multiple sequencing-based methods that can be used to analyse cfDNA. The techniques differ in the way the library is prepared, the size of the genome that will be analysed, and the mean sequencing depth per sample. In the following sections, I will compare different methods for cfDNA library preparation and sequencing. I will also highlight the advantages and disadvantages associated with each.

1.3.1 Preparing next-generation sequencing libraries from cfDNA

Multiple strategies are currently available to prepare a DNA sample for next-generation sequencing. Both digital polymerase chain reaction (dPCR) and library preparation methods are commonly used in ctDNA analysis; however, this chapter will mostly focus on library preparation methods. A library preparation for cfDNA usually entails end repair, appending sequencer-specific ‘adapters’ to the sequence of interest, and amplification of the fragments (see Fig. 1.3) [37–39]. This final step is necessary due to the low concentrations of input

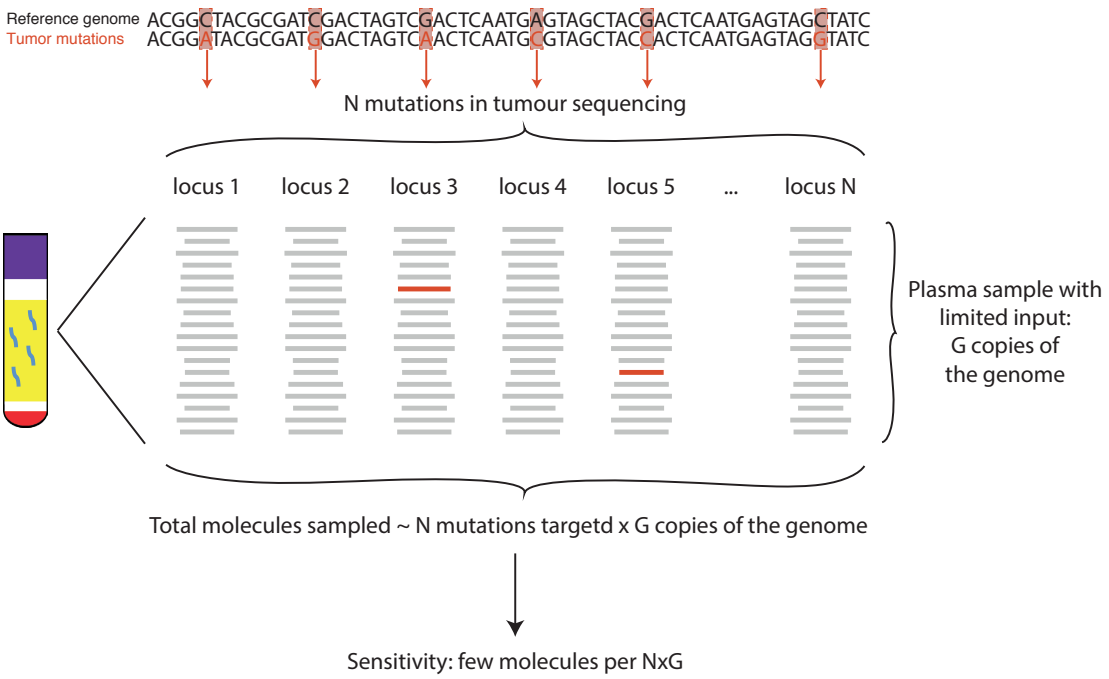


Fig. 1.2 **Benefit of multi loci sampling in ctDNA detection.**
A given sample contains a limited number of copies of the genome, denoted by G. For plasma samples the small amount of material, generally on the order of a few thousand genomes (also referred to as copies), limits the sensitivity that is attainable in detection of individual mutations to one mutant copy per G copies present in the sample (e.g. 1 copy per 10,000), even if the assay employed has low background noise and can provide better resolution (e.g. below $1x10e^{-5}$). By analysing in parallel a large number of marker loci (e.g. loci that are found to be mutated in the patient's tumour), denoted by N, detection of tumour DNA can be substantially enhanced to detect one or few mutant molecules per NxG copies. For example, with N=1000 mutations and G=3000 copies, this provides a total of 3,000,000 reads for analysis and can achieve detection to parts per million and beyond. The same approach can be employed for other applications which aim to detect non-background/altered DNA, such as detection of fetal DNA or DNA from transplanted organs, in limited amounts of material such as plasma samples or other body fluids.

DNA that are commonly available from cfDNA [38, 40]. While tissue or germline DNA library preparation kits require a fragmentation step to generate short fragments (∼200bp) compatible with the sequencer (a potential source of severe fragment loss [41]), cfDNA samples do not have to undergo DNA shearing due to its already fragmented nature. One of the main challenges when sequencing DNA libraries is to mitigate the bias introduced by PCR amplification [39]. This challenge is particularly important in cancer genomics when attempting to sensitively detect SNVs targetable by precision medicine [42]. Another consideration is the maintenance of molecular complexity during library preparation to increase the chances of finding rare mutations with low allelic frequencies. Current methods typically vary in the efficiency with which the final output represents the molecules which went in to the reaction [39, 43]. The two most common library preparation techniques in the cfDNA field (double and single stranded library preparation) are described in more detail in the next section.

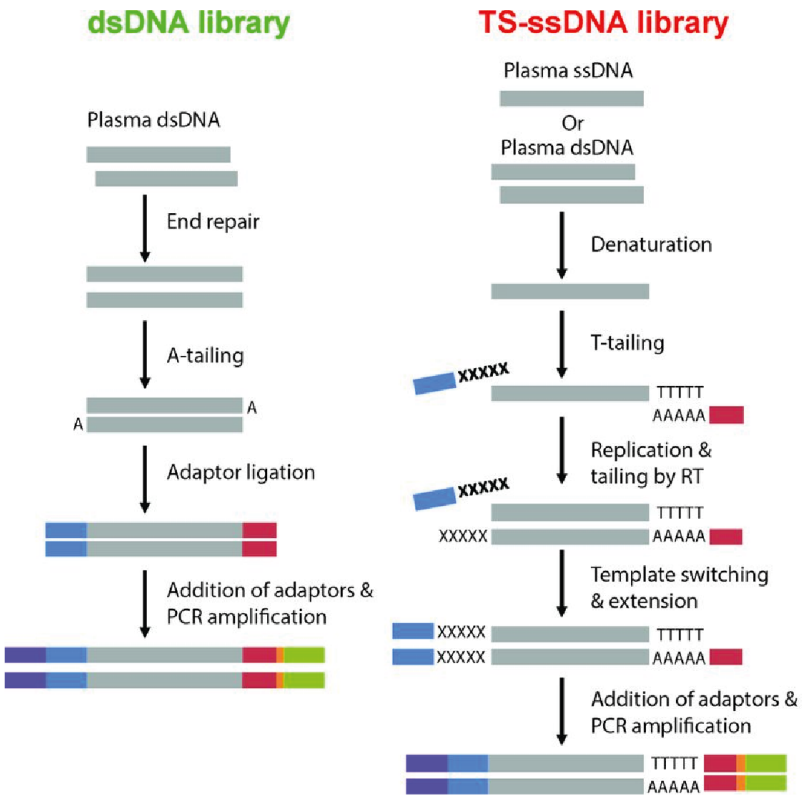


Fig. 1.3 **cfDNA library preparation methods.**
Comparison of double stranded and single stranded library preparation methods for cfDNA. In general samples undergo end repair, appending of sequencer-specific 'adapters' to the sequence of interest, and amplification of the fragments. Unique molecular identifiers can be added for additional error collapsing. Adapted from Vong et al. [44]

1.3.2 Library preparation of double-stranded DNA

The most commonly used library preparation protocols in the cfDNA field are aimed at sequencing double-stranded DNA. Indeed, a multitude of kits and methods are available that follow a similar workflow. The cfDNA sample first undergoes end-repair, followed by a phosphorylation of the 5' ends. The 3' ends undergo an A-tailing step to allow for ligation of the sequencer-specific adapters. Upon adapter ligation, samples undergo a few cycles of PCR to enrich the fragments that are compatible with the sequencer (Fig. 1.3). This is followed by a final 'clean up' of the library to remove excess adapter dimers and hetero-duplexes [39–41]. Individual DNA samples can be tagged with uniquely barcoded adapters. These barcodes consist of a short and unique stretch of bases and allow for multiplexed sequencing. Molecules originating from the same sample will obtain the same string of unique bases, therefore, each barcoded-DNA fragment in the sequencing pool can be attributed to its sample of origin [42].

Competition in the market of library preparation kits has helped to lower the prices of the kits while also improving their quality. The required input for most kits starts at less than a nanogram of DNA, allowing for sample preparation even when very little material is available [43].

The most commonly used double-stranded library preparation protocols tend to capture fragments greater than 100bp in length [45]. It is still unclear why shorter fragments are not observed but there are different hypotheses. Shorter molecules could be lost during the extraction or library clean up step, be damaged or be of single stranded nature and thereby not captured by the double-stranded protocol. This results in the loss of fragments shorter than 100bp, which may be enriched for ctDNA [14, 15]. Single-stranded library preparation methods have emerged as an alternative library preparation approach that may overcome this problem [45, 46].

1.3.3 Single-stranded DNA library preparation

Developed for the analysis of ancient, degraded DNA, current single-stranded DNA library preparation methods allow for the capture of single-stranded, double-stranded, and damaged DNA fragments [46]. Therefore, as compared to double-stranded library preparation protocols that capture only double-stranded DNA, single-stranded DNA library preparation allows for the capture of a broader range of DNA fragments (Fig. 1.3). Indeed, in a study that used a single-stranded protocol on DNA from a cohort of transplant patients, a greater portion of 50 - 100bp fragments was recovered using this approach as compared to a double-stranded equivalent [45]. Also, when applied to plasma samples from healthy individuals, a greater

representation of shorter DNA fragments was recovered as compared to the standard double-stranded library preparation [47]. An updated version of the Gansauge and Meyer ancient DNA protocol was published in 2017, showing greater recovery of DNA, and an improved turn-around time [48]. Previous research suggested an enrichment of ctDNA in shorter fragments [14, 15, 49, 50]. Therefore, the use of single-stranded DNA library preparation protocols, that improve the coverage of shorter fragments, might be expected to improve the recovery, and detection of, ctDNA. However, initial reports seem to indicate that there is no improvement in ctDNA detection using a single-stranded DNA library [51], raising questions as to the origin of the additional DNA fragments that are captured by this protocol.

Other single-stranded protocols include the commercially available SMART ChIP-Seq kit that is based on template switching. This protocol was initially developed for RNA approaches [52, 53]. Compared to the ancient DNA protocols described above, the SMART ChIP – Seq protocol involves a simpler, and quicker, workflow. However, a recent publication indicates that this method preferentially captures fragments with poly dA/dT tracts, which could induce a bias in cfDNA recovery [54]. Similar comparisons should be performed on the other library preparation methods to obtain a clearer picture of the sequencing biases specific to each method.

Here I have introduced different methods to prepare sequencing libraries from cfDNA samples. Their main advantages and disadvantages are highlighted in table 1.1. One of the most commonly used approaches in the field is double-stranded DNA library preparation. There are a variety of high quality and easy to use kits available, making it a robust method for the preparation of cfDNA samples for sequencing. However, double-stranded library preparations will capture mostly non-degraded fragments longer than 100bp, thereby not representing all cfDNA populations. Single-stranded DNA library preparation methods have partly addressed this problem by allowing the capture of both shorter and degraded DNA fragments, as well as single- and double-stranded DNA.

1.3.4 Enrichment of circulating tumour DNA using size selection

Single-stranded library preparation protocols aid in capturing shorter DNA fragments, which might be biased towards a higher proportion of ctDNA, but do not actually enrich for ctDNA. One means by which one could potentially do so is by carrying out size-selection. As early as 2010 it was observed that ctDNA fragments are predominantly shorter than 150bp, while cfDNA of non-cancerous origin is predominantly 166bp in length (Fig. 1.4) [9, 14, 15, 27, 50, 55, 56]. Therefore, focusing the analysis on the shorter fragment lengths could enrich the sample for ctDNA over cfDNA, making its detection easier. In 2011, Mouliere and colleagues showed an enrichment for ctDNA relative to cfDNA at the shorter fragment

lengths in xenograft samples [14]. In 2016, Underhill and colleagues applied this concept to human samples and could show that mutant alleles of *BRAF* V600E in melanoma patients were more commonly observed at shorter fragment lengths compared to wild-type fragments. Similarly, an enrichment for *EGFR* T790M was observed in lung cancer patients when selecting for shorter cfDNA fragments [15]. Mouliere and colleagues confirmed the ctDNA enrichment in shorter fragment lengths across 344 plasma samples from 200 cancer patients varying in cancer type and stage [27]. Selecting for the region of 90 - 150bp, the study shows a more than two-fold median enrichment for ctDNA across all samples after size selection.

Both, *in silico* and *in vitro* methods have been used to select for the shorter fragments [15, 27]. *In silico* methods select for the fragments of interest only after sequencing, allowing the researcher to apply different size selection windows to the same sample. This is especially useful if the same sample is used for additional analysis not requiring size selection or if the best enrichment window is not previously known. However, this approach requires additional sequencing to ensure sufficient reads are remaining for downstream analysis after selecting for the fragment length region of choice, resulting in higher sequencing costs. *In vitro* methods on the other hand select for the region of interest before the sample undergoes sequencing, usually using a form of gel-based selection [15, 27]. Only the region of interest is retained and sequenced, reducing the total required sequencing of the sample as compared to the *in silico* selection. While this approach could be cheaper overall, the main drawback is the sample loss in the process. Not only are any fragments outside the window of interest lost for downstream analysis but even within the selection window fragment selection is not 100% efficient.

1.3.5 Sequencing error correction using unique molecular identifiers

While selecting for certain fragments may lead to more confident mutation detection, the general level of background noise remains a challenge in the field. Indeed, with an error rate of just below 1%, current sequencing methods yield data with high background noise [41, 57]. Additionally, given the low concentration of cfDNA, PCR amplification is usually warranted during library preparation. This further leads to the induction of PCR errors that may reach allelic fractions as high or higher than true mutations [58, 59]. Methods have evolved to aid in the downstream analysis and differentiation of true mutations from noise.

The concept of uniquely tagging individual molecules was first described in 2003 in the context of determining the unique number of mRNA molecules in a given sample [60]. In 2011, unique molecular identifiers (UMIs) were first used to reduce PCR background noise and obtain a cleaner sequencing result [61, 62]. UMIs are composed of a string of entirely random nucleotides that tag each molecule of the initial sample in a unique way [61].

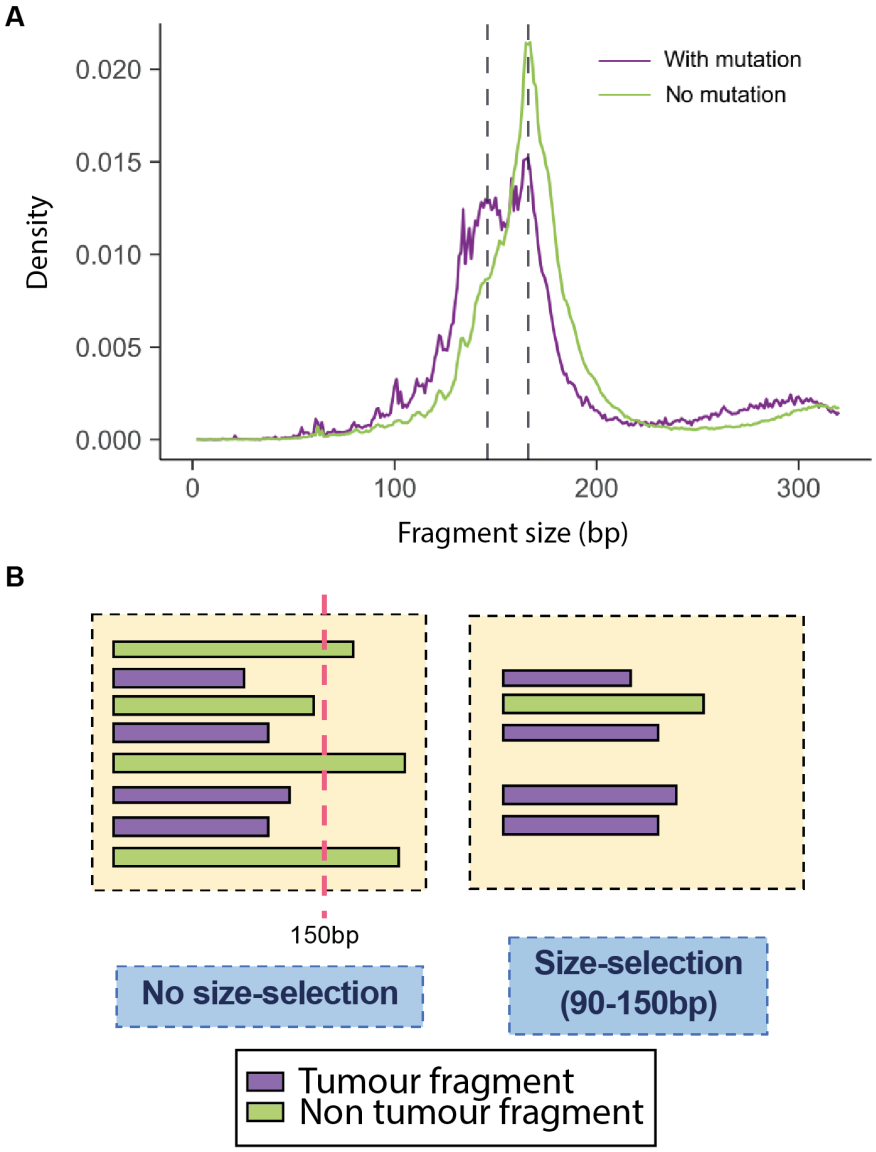


Fig. 1.4 **Potential for size selection on cfDNA.**
(**A**) ctDNA tends to show a shorter fragment length than non cancerous cfDNA, allowing for sample enrichment by selecting shorter fragment length. Adapted from Mouliere et al. [27]. (**B**) A given cfDNA sample can be enriched for tumour derived ctDNA fragments by size selecting the sample for fragments shorter than 150bp. The generally shorter ctDNA fragments (shown in purple) will mostly be retained while the longer cfDNA fragments (shown in green) will be removed. Figure adapted from an initial figure prepared by Irena Hudecova.

After sequencing, the UMIs are used to group fragments of the same origin into families which are then combined into a single consensus sequence [63]. As shown in Fig. 1.5, this consensus sequence aids in identifying PCR and sequencing errors, which should only be present in some, but not all, members of a given family. Conversely, true mutations should be represented in most, if not all, of the members of that family. Thus, the use of UMIs allows for error suppression and more stringent and confident calling, which improves downstream analyses. Collapsing reads purely based on start- and end-position is not advised since the size distribution of cfDNA is too tight and different unique molecules could have the same start and end position by chance [64].

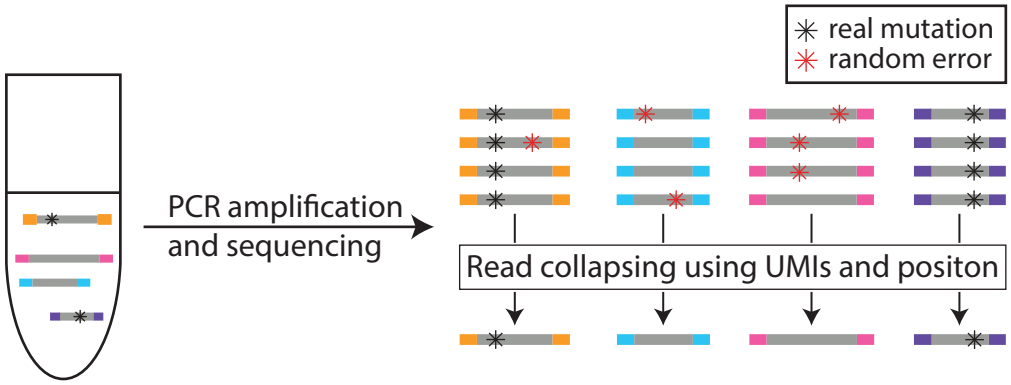


Fig. 1.5 **Schematic representation of background noise reduction through UMIs.** Unique molecules are tagged with different adapters before undergoing PCR amplification and sequencing. Utilising read position and UMI sequence information, PCR and sequencing errors are removed and molecules are collapsed into a consensus sequence.

UMIs have started to make an impact within the field of ctDNA research. Recent examples of their application include the Safe–SeqS approach with an error rate of only $3.5\text{x}10^{-6}$ mutations/bp (70-fold less than without using UMIs), which was used for the multi-cancer study by Bettegowda and colleagues [30, 58]. UMIs were also incorporated into the iDES CAPP-Seq approach used by Newman and colleagues to study early-stage lung cancers. Using this approach, the authors showed that, together with in-silico background polishing, a sensitivity of 0.0025% could be achieved [65].

Despite their promise, there are some limitations when using UMIs. For example, it is possible that PCR errors that occur early in the amplification process may not be filtered out. Furthermore, sequencing errors in the sequence of the UMI itself may lead to assignment of fragments to the wrong family [57].

The widespread development of high-depth sequencing for ctDNA mutation analysis, as well as the availability of standardised kits, lead to an easier implementation of the approach without the necessity of prior experience [66]. Tools such as CONNOR, MAGERI, UMI-tools, and Agilent's SureCall further aid with the analysis of UMI data [63, 67–69].

Building upon the background noise suppression achieved through the use of UMIs, Schmitt and colleagues developed a technique called duplex sequencing. They not only incorporate UMIs into their library preparation but also retain strand directionality and origin to further reduce background noise [70]. After generating a strand-based consensus sequence, they identify which two strands came from the same original double-stranded molecule and combine those two consensus strands into a final duplex consensus sequence. They estimate their error rate to be at 3.8×10^{-10} or even lower, allowing for very sensitive mutation calling. However, this degree of read combining warrants a great initial sequencing depth, resulting in a much increased sequencing cost [70].

Indexing with unique molecular barcodes has been proposed to counteract potential PCR and sequencing biases that arise during the standard preparation process. While indexing will remove a great number of PCR and sequencing errors, they might still fail to correct for early PCR errors or lose data due to PCR and sequencing errors in the barcodes themselves.

A summary of the main options and new developments available for cfDNA library preparation, ctDNA enrichment, and error rate reduction are compared in table 1.1. I will now turn towards the different sequencing methods available for cfDNA analysis.

Method	**Main advantages**	**Main disadvantages**
Double-stranded library preparation	Quick and optimised protocols Low input	Loss of shorter and degraded fragments
Single-stranded library preparation	Capturing short and degraded fragments Low input	Bias in fragment recovery More complex protocols
Unique molecular identifiers	Reduced noise by read collapsing into families	Greater sequencing depth required Unable to identify early PCR errors UMI sequencing errors result in wrong collapsing

Table 1.1 **Comparison of library preparation methods.** Highlighted are the major advantages and disadvantages of the library preparation methods discussed in this chapter.

1.3.6 Tailoring approaches for cell free DNA sequencing

Upon generation of a library, different sequencing methods are available depending on the required analysis (Fig. 1.6). One can either proceed directly with whole genome sequencing, or enrich the sample for some part of the genome. In the following sections, different sequencing approaches will be explained in more detail and their advantages and disadvantages, as well as their application to the study of plasma DNA, will be discussed.

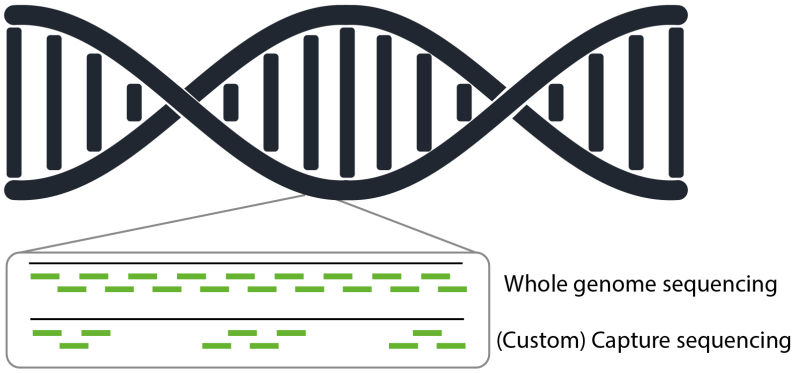


Fig. 1.6 **Sequencing methods for cfDNA analysis.**
cfDNA can be analysed by targeted and untargeted sequencing approaches. Whole genome sequencing will cover the entire genome while capture based approaches will focus on a subpart of the genome. This can be accomplished using a commercial kit covering (for example) the entire exome (whole exome sequencing) or by designing a custom capture sequencing panel covering specific regions or genes of interest.

Whole genome sequencing

Whole genome sequencing (WGS) of plasma DNA provides researchers with a comprehensive view of the entire genome of the patient [71], facilitating the analysis of somatic copy number alterations (CNAs), structural variants (SVs), and SNVs. With the development of digital karyotyping, it has become possible to analyse copy number changes of the genome in greater detail, even with sufficient sensitivity to observe focal events involving small genomic regions (<1Mbp) [72]. Digital karyotyping laid the foundation for later copy number analysis and its implementation to plasma analysis. Whole genome sequencing analysis of cancerous plasma samples was first employed in 2012 when CNAs were identified, even in the absence of matched tumour specimens [73]. Using CNAs to detect cancer represents a useful tool since amplifications and deletions are often an early event in cancer development [74]. Additionally, it is possible to infer the copy number status from low-depth data,

even with a sequencing depth of <1x, in turn allowing simple and cost-effective interrogation of ctDNA [75, 76]. The ability to detect CNAs even at low depth has led to the use of terms such as shallow whole genome sequencing (sWGS) or low-pass whole genome sequencing.

For sWGS-based CNA analysis, the genome is apportioned into 'bins' of an equally sized length (ranging from kb-Mb). The sequencing reads that map within each of these bins are counted and this value is compared to the average bin read count across the entire genome. Any bins spanning regions that are amplified or deleted will contain greater or fewer reads than the respective average bin count. Such a ratio is obtained for every bin across the length of the genome, allowing one to generate a CNA profile. This graphic representation of the genome allows to quickly visualise amplifications and deletions, as well as compare the overall pattern between different samples from the same patient, or between different patients. The chosen bin size is correlated with the sequencing depth where greater sequencing depth allows for a smaller bin size and, therefore, a more detailed CNA profile.

One can convert the above CNA profiles into quantifiable metrics to estimate ctDNA levels, for example, using a z-score as described by Heitzer and colleagues [77]. This method determines a score for each sample based on the deviation in read distribution from a cohort of control plasma samples [77]. A sequencing depth coverage from 0.1x to 0.2x is sufficient to determine the z-score of a patient [77]. Similar metrics such as ichorCNA [75] and t-MAD [27] have since been published, all allowing quantification of copy number alterations.

One can also identify SNVs from WGS data; however, greater sequencing depth is required to ensure sufficient confidence in mutation calling. Indeed, the detection limit of SNVs is directly correlated with the coverage; the more a sample is sequenced, the lower the mutation frequency detection threshold (i.e. greater sensitivity). At a sequencing depth of 17x, Chan and colleagues identified both CNAs and SNVs in plasma samples from four cancer patients [17]. Comparing the copy number data from plasma and matched tumour tissue, the authors observed a better correlation with increasing tumour size and ctDNA fraction in the plasma [17]. For one patient with two different cancer types, CNA analysis suggested that the copy number alterations of both cancers could be detected in the plasma, demonstrating that plasma analysis has the potential to overcome spatial tumour heterogeneity [17].

When analysing SNVs, the authors could show that mutations shared between different tumour regions in the same patient were better represented in the plasma, as reflected by their higher allele fraction [17]. Conversely, sub-clonal tumour mutations were less well-represented and had a greater chance of being missed [17, 75].

In summary, whole genome sequencing is a powerful tool to obtain genome-wide information for a sample but becomes increasingly cost prohibitive as the required depth of

sequencing increases. Indeed, even a modest overall depth of ∼30x (generally used to genotype polymorphisms) typically requires a high enough sequencing cost such as to preclude use in most clinical and research settings.

Capture-based sequencing approaches

While WGS of plasma DNA can provide a comprehensive overview of a patient's genome, it lacks sensitivity for the detection of low-frequency mutations due to the current cost of sequencing. By restricting the size of the region to be analysed, one can reach a greater depth per sample while maintaining or decreasing the cost of sequencing. This greater depth results in a greater sensitivity and the ability to reliably detect mutations at lower frequencies [71]. After library preparation, samples can undergo a 'hybrid capture' process. In this method, regions of interest are selected by the annealing of complementary DNA or RNA 'baits' or primer pairs. Captured regions are amplified while uncaptured regions are washed away. Depending on the design of the capture baits, the captured regions may include the entire exome (whole exome sequencing, WES), or some other custom-designed region (custom capture sequencing). In the case of the latter, current approaches typically target commonly mutated genes, or already identified mutations. A major limitation of capture-based approaches can be uneven coverage, due to the fact that some regions lack complexity and are difficult to target. As a result, they will not be captured and enriched during the hybridisation [78].

Whole exome sequencing

Assuming that most driver mutations will occur in the actively transcribed part of the genome, targeting the exome allows researchers to focus their sequencing efforts on this ∼1% of the genome. Many companies offer exome capture reagents, varying in the capture approach, complexity of the protocol, and covered regions [79].

WES has been demonstrated to be a useful tool for monitoring tumour evolution. Murtaza and colleagues applied WES to serial plasma samples with high (>10%) levels of ctDNA. An input of 2.3ng DNA (690 genome equivalents) and a sequencing depth of 31-160x was sufficient to call mutations from patient plasma [18]. Using longitudinally obtained samples, it was possible to track tumour evolution throughout treatment [18].

Furthermore, Girotti and colleagues applied WES to plasma samples from stage III melanoma patients. They successfully identify resistance mutations, highlighting the potential for ctDNA as a disease monitoring tool [80]. Dietz and colleagues showed the utility of WES on low volume serum samples from stage III NSCLC patients. However, due to limited

sample input and sequencing depth (68x) they validated only 17% of mutations identified in matched tumour specimens [81]. Nevertheless, they also identified a potential resistance mutation that was absent in the corresponding tumour sample, highlighting how ctDNA can overcome spacial tumour heterogeneity [81]. WES was applied to plasma samples from metastatic cancer patients in a study by Butler and colleagues [82]. They showed good correlation between tumour and plasma mutations and again identified additional plasma mutations that were absent in the tumour sample [82]. However, their study focused on only two patients, utilising 15mL and 25mL of plasma at sequencing depths of 309x and 561x respectively [82].

While there is a growing number of studies showing the feasibility of WES on plasma and serum samples, it is important to consider the current limitations. All studies thus far have focused on late stage cancer patients, in whom levels of ctDNA are generally higher [30, 64]. Additionally, the sequencing depth in these studies is quite high which, whilst allowing for greater sensitivity in mutation calling, increases the overall cost of this approach. Applied to a larger cohort or in a clinical setting, WES would generally not be currently feasible.

Custom capture sequencing

As an alternative to 'off the shelf' exome capture kits, custom capture methods have emerged, allowing one to focus sequencing efforts on particular genomic loci of interest. This, in turn, means that either the same number of samples can be sequenced to a greater depth, allowing for more sensitive mutation calling, or more samples can be sequenced to the same depth, enabling application to larger cohorts.

Cancer Personalized Profiling by Deep Sequencing (CAPP – Seq) is one example of a custom capture sequencing approach that relies on the presence of recurring mutations across a cancer cohort. Using publicly available WES and structural rearrangement data, common mutations and rearrangements are identified and used to design a targeted gene panel [31]. Focusing on NSCLC, a 125kb panel was designed and applied to plasma samples from stage I – IV patients. The small panel size allowed for deep sequencing ($\sim$10,000x in this study) at a reasonable cost per sample. For late stage patients, CAPP – Seq showed 100% sensitivity for ctDNA detection and proved advantageous over imaging technology for disease detection, emphasising the potential of capture-based methods for cancer diagnosis and monitoring [31]. The same authors further developed the iDES CAPP – Seq approach described previously (see section 1.3.5), a technique that implements molecular barcodes and background polishing of sequencing 'noise'. These developments allow for a sensitivity of 0.0025%, sufficient for the detection of ctDNA at very low levels [30, 65].

Phallen and colleagues also used a targeted capture approach to detect ctDNA in patients with different cancer types and stages of disease. Their targeted error correction sequencing (TEC – Seq) method utilises an 81kb gene panel containing 58 cancer related genes. Error suppression of their data is based on the start and end position of the reads, as well as a small number of dual-index barcode adapters [64]. Using their general gene panel and deep sequencing with 30,000x coverage, Phallen and colleagues were able to detect ctDNA in 62% of all stage I and II patients and 77% of all stage III and IV patients [64]. Similar to the analysis of Bettegowda and colleagues, they also found a difference in detection between different cancer types and an increased detectability with later stage disease [30, 64].

Alternative sequencing based methods

Beyond hybrid capture based approaches, other methods allow for sensitive mutation detection. For example, tagged amplicon deep sequencing (TAm – Seq) is an amplification-based approach that combines singleplex and multiplex amplification steps [35]. Regions of interest are first enriched by amplification in a multiplexed PCR, thereby reducing later sampling bias. The sample is then split and a singleplex PCR of each of the regions of interest is carried out [35]. Amplicons are selected based on publicly available data or cohort-specific knowledge of mutations [35]. The size of the amplicons is around 100bp, accounting for the majority of cfDNA fragments based on the known distribution of fragment lengths [9, 55]. TAm – Seq has a similar sensitivity to dPCR, and reduces sampling bias through the first multiplexed amplification step [83].

The recent TRACERx study used patient-specific amplicon sequencing panels for the enrichment of libraries to improve detection and sensitivity [19]. Based on WES tumour data, a median of 18 SNVs were identified for each of 96 patients. These regions were enriched in libraries prepared from plasma samples. While not as sensitive as iDES CAPP – Seq, the approach used by Abbosh and colleagues was able to detect ctDNA in 94% of stage I squamous cell carcinoma NSCLC patients. On the other hand, the ctDNA detection rate in patients with adenocarcinoma NSCLC was only 12.8% at stage I, highlighting the interaction between ctDNA detection and cancer (sub-)type [19, 30].

When choosing the most appropriate sequencing method, it is important to have a good estimation of the expected levels of ctDNA in a given sample. Depending on the ctDNA fraction, more or less sensitive methods will be required for analysis. One potentially useful practice could be to implement sWGS of a given plasma sample prior to choosing the analysis approach. This is because sWGS is a comparatively cheap and rapid technique that can assess likely ctDNA levels. Indeed, the mFast-SeqS and ichorCNA approaches have been demonstrated to aid in determining ctDNA concentration in the plasma, in-turn identifying

the best analysis platform [75, 76]. These approaches allow the user to group the samples into low and high tumour burden ctDNA samples and, based on this classification, one can then choose an approach for the analysis of the specific alterations in the sample. The only caveat to this would be the reliance on somatic CNAs – this approach would not be as effective in cancers known to carry few somatic CNAs [84].

I have presented in this section various methods for sequencing cfDNA from plasma samples. Fig. 1.7 provides a schematic overview of the preparation process involved in the described methods. The main difference between these methods is the size of the genome that will be sequenced and the resulting sensitivity, specificity, and sequencing cost per sample. The methods also differ in the type of somatic event they can detect. While WGS provides the broadest and least biased overview of a given sample, it is also the most expensive method if an appreciable depth of sequencing is required. Using capture-based methods, one decreases the window of the genome that is being targeted, thereby reducing the total sequencing cost per sample. Off-the-shelf capture approaches exist that target the whole exome or a cancer specific gene lists. Alternatively, one can design custom capture panels to improve the sequencing depth of bespoke regions of interest while maintaining the same overall cost. Lastly, alternative sequencing approaches such as amplicon sequencing were presented which can also be used for sensitive detection of ctDNA.

1.4 Lung cancer

Lung cancer is the second most abundant cancer in both, males and females. It is more common in older age groups and tends to be slightly more common in males than in females [85]. The single most common cause of lung cancer (close to 90% of all cases) is smoking [86]. The incidence of lung cancer is related to both, the years of smoking and the number of cigarettes per day. Other common risk factors include a family history in lung cancer, exposure to arsenic or asbestos, and living in areas with strong air pollution [87]. Lung cancers from smokers tend to show a larger transversion rate and higher mutation burden compared to lung tumours of non-smokers [88]. Especially the cytosine to adenine transversion is associated with tobacco smoking is predominantly seen in adenocarcinomas of smokers [89].

In 2012, over 44,000 new cases of lung cancer were diagnosed in the UK and accounted for 13% of all new cancer cases that year, making it the third most common cancer [90]. Due to a decrease in smoking prevalence, lung cancer incidence rates have decreased since the early 1990s. However, the incidence rate in females is still increasing, due to their later

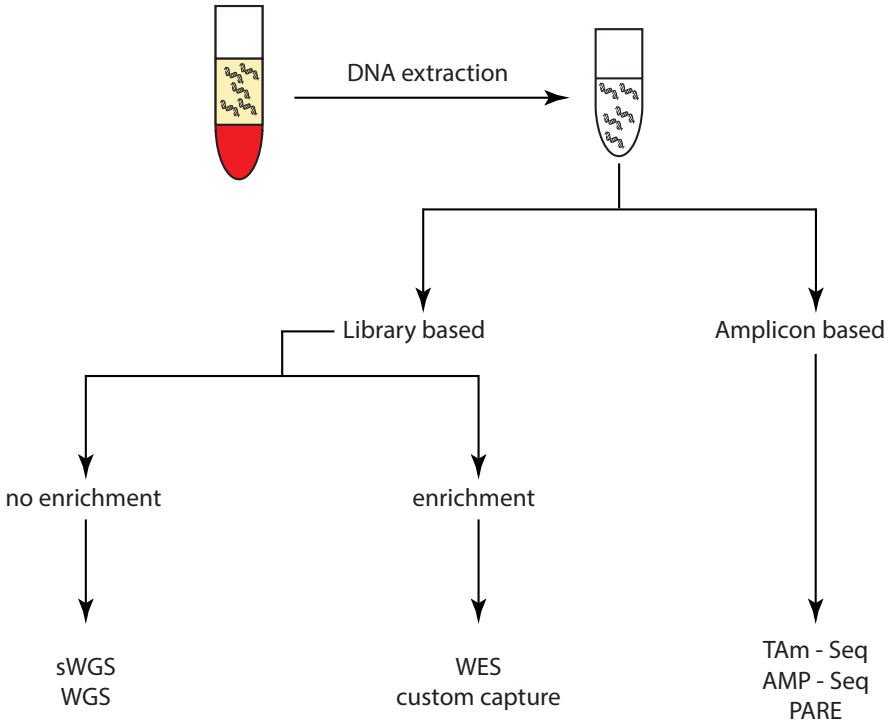


Fig. 1.7 **Overview of preparation process of different cfDNA sequencing methods described in this chapter.**

uptake of smoking [87, 90]. Since the 1990s, incidence in males has decreased by 31%, while it increased by 28% in females in the UK [90].

The five main histological subtypes of lung cancer are adenocarcinoma, squamous-cell carcinoma, small-cell carcinoma, large-cell neuroendocrine carcinoma, and pulmonary carcinoid tumours, with adenocarcinoma, squamous-cell carcinoma, and small-cell carcinoma being the most common [89, 91]. The majority of the genomic alterations found in the various subtypes are specific to the given subtype and only few events are shared between subtypes [89]. During my PhD I was mostly working on patients presenting with non-small cell lung cancer (NSCLC), which makes up 85 – 90% of all lung cancer cases. The remaining cases are mostly small cell lung cancer [92]. NSCLC grows and spreads more slowly than small cell lung cancer and contains three main subclasses, depending on the tumour morphology (40% adenocarcinoma, 25 – 30% squamous cell carcinoma, and 10 – 15% large cell carcinoma) [93].

In the UK, lung cancer is mostly diagnosed at the later stages. In 2014, 67% of the diagnosed cases were stage III or IV [90]:

	England	**Scotland**	**Northern Ireland**
Stage of Diagnosis	**2014**	**2014-2015**	**2010-2014**
Stage I	15	18	13
Stage II	7	9	8
Stage III	19	22	22
Stage IV	48	46	43
Stage Unknown	10	5	15

Table 1.2 **Stage at diagnosis of lung cancer in the UK.** Percentage of cases diagnosed at the respective stages (%). Data taken from Cancer Research UK [90]

Lung cancer is the leading cause of cancer death, being responsible for more than 20% of all deaths associated with cancer [94]. The survival of lung cancer is extremely low with only 5% of patients still being alive 10 years after their diagnosis [90]. Over the last 40 years there have only been marginal improvements in the five and ten year survival rates [90]. However, patients are now twice as likely to survive for the first year after their diagnosis (table 1.3).

The survival of lung cancer is highest for patients diagnosed at earlier stages. As seen in Table 1.4, 35% of patients diagnosed with stage I disease will still be alive after five years, while only 6% of stage III patients will still be alive (see Table 1.4). For patients diagnosed with stage IV, survival rates past two years are too low to even be calculated [90].

Period of Diagnosis	1 year	5 year	10 year
1971-1972	15.9	4.6	3.1
1980-1981	18.4	5.6	3.8
1990-1991	20.4	6	3.9
2000-2001	24.2	6.8	3.9
2005-2006	27.8	8	4.3
2010-2011	32.1	9.5	4.9

Table 1.3 **Lung cancer survival rates over time.** Survival is shown in precent. Data was taken from Cancer Research UK [90]

Stage at Diagnosis	5 year survival (%)
Stage I	35.33%
Stage II	20.89%
Stage III	6.32%
Stage IV	unavailable
Stage Not Known	5.79%
All Stages	9.68%

Table 1.4 **Lung cancer 5 year survival by stage.** Survival is reported in percent. Data was taken from Cancer Research UK [90]

Even though the abundance of lung cancer is quite high, little improvement has been made with regards to overall survival. After 40 years of active research, the long-term survival after being diagnosed with lung cancer is still amongst the worst of all cancers [95]. This can be attributed to lung cancer being mostly diagnosed at the later stage of disease. The symptoms for lung cancer are unfortunately not specific for the disease but also indicative of other respiratory diseases, They include chest pain, cough, difficulty breathing, fatigue, chest infection and weight loss [96, 97]. As a result, lung cancer is often only diagnosed at the advanced stages, considerably reducing the overall survival chances of the patient.

1.4.1 ctDNA in non-small cell lung cancer

Liquid biopsies have been proposed as an alternative of cancer detection and in some cases have been approved for the clinic [16]. In lung cancer, ctDNA has shown sensitive detection in the later stages but detection in the early stages has been limited [28]. A recent review compared the currently published literature on ctDNA detection in early stage NSCLC [28]. Four different studies [19, 36, 64, 98] were highlighted with regards to their detection rates for different tumour stages as well as the ctDNA levels in the detected cases. The studies utilise either gene panels and hybridisation technology [36, 64] or multiplex PCR [19, 98] to detect ctDNA. The CAPP-Seq assay and TRACERx study were aimed at the detection of MRD, utilising information on the mutations seen in the tumour, while the CancerSEEK and TEC-Seq assay took an untargeted approach geared towards the detection of cancer [28]. Even though all four studies analysed patients with NSCLC, there was some variability in the patient cohorts of the studies with regards to stage and tumour subtype. For example, the cohort used for the CAPP-Seq study mostly contained late stage patients, while the cohort used for the TEC-Seq application was biased towards early stage patients. The TRACERx and CancerSEEK studies showed an even representation of all cancer stages (see Table 1.5).

Study	**I**	**II**	**III**	**IV**	**Total**
CAPP-Seq	5 (15.6%)	6 (18.8%)	21 (65.6%)	0 (0%)	32
TEC-Seq	29 (41.4%)	31 (44.3%)	5 (7.1%)	5 (7.1%)	70
CancerSEEK	46 (44.7%)	26 (25.2%)	31 (30.1%)	0 (0%)	103
TRACERx	59 (61.5%)	23 (24.0%)	14 (14.6%)	0 (0%)	96

Table 1.5 **NSCLC patients by stage for different studies.** Data is adapted from Abbosh et al 2018 [28]. The CancerSEEK assay is utilising both ctDNA and protein markers to achieve a more sensitive detection.

With regards to cancer subtype, the CAPP-Seq study had a balanced representation of the different subtypes [36], the TRACERx and TEC-Seq study were biased towards adenocarcinomas [19, 64] and the CancerSEEK study did not provide information on tumour subtype [98]. These differences with respect to tumour subtype make it more challenging to compare directly between the four studies and their respective detection rates (see Table 1.6). Due to these varieties in study design and the constellation of the respective cohorts, the results from these four studies cannot be compared directly but rather provide an overview of the ctDNA levels in NSCLC.

Study	**Stage I**	**Stage II**	**Stage III**
TRACERx	37.29%	69.57%	57.143%
CAPP-Seq	100%	66.67%	95.24%
TEC-Seq	44.83%	74.19%	80.00%
CancerSEEK ctDNA only	4.35%	38.46%	35.48%
CancerSEEK ctDNA + proteins	43.48%	69.23%	74.19%

Table 1.6 **ctDNA detection rates by stage for different NSCLC studies.** Data is adapted from Abbosh et al 2018 [28]. For the CancerSEEK assay detection is reported when just relying on the ctDNA based assay ("ctDNA only") or when using ctDNA together with the protein biomarker parameters ("ctDNA + proteins"), resulting in a more sensitive detection.

As part of the review, Abbosh et al analysed the maximum detected ctDNA fractions per patient. They found an increasing maximum ctDNA fraction with increasing stage, consistent with previous pan-cancer studies [28, 30]. The median maximum mutant allele fraction observed across the studies for patients with stage I, II, and III disease were 0.31%, 0.48% and 1.48%, respectively [28]. Similar to other reports [30], the detection rates of ctDNA were lowest for stage I disease, with the exception of the CAPP-Seq assay, showing perfect detection in the small cohort of stage I patients (n=5) [28]. The low detection rates in stage I disease are thought to either be due to tumours not releasing ctDNA or releasing it in such low concentrations that its detection becomes impossible given the current analysis platforms [28].

ctDNA detection and monitoring has been used in cohorts of early stage NSCLC patients with limited success. While detection rates look promising in patients with stage II and III disease, detection of ctDNA in stage I patients has been inferior. It is currently unclear if this decreased sensitivity is related to the smaller tumour size in early stage disease, resulting in limited ctDNA release, or if additional biological factors prohibit its release. In case of the former, development of a more sensitive ctDNA detection tool should improve detection rates in this challenging setting.

1.5 Limiting sample input in cfDNA analysis

As described in section 1.3.6, CNV profiles and other metrics can be used to detect ctDNA from very low depth sequencing (usually 0.1x). If sequencing is performed to such low depth, only a small proportion of the genetic diversity in a given sample is analysed, raising the question if the same sensitivity could be achieved from samples with limited sample volume. If this were possible, it could allow for even more frequent sampling of the patient as smaller volumes are easier to obtain and the chance of successful sample collection should be higher. In theory, a single drop of blood with a volume of 50µL should contain sufficient amounts of cfDNA for analysis by sWGS. Given the general concentration of cfDNA (5-10ng/mL), a single droplet of blood should contain 0.25-0.5ng of DNA. 1ng of DNA is roughly equivalent to 300 genomic copies, meaning that a single droplet of blood could contain 75-150 genomic copies of cfDNA, which should be more than enough to be analysed by sWGS.

The main advantage when using blood spots is the ease of sample collection and processing. Without the need for specifically trained research staff, samples could even be collected at home and sent to the lab for analysis. Real-time PCR has previously been used to carry out fetal RHD genotyping and HIV detection using maternal dried blood spots [99, 100]. A pilot study in breast cancer patients performed whole genome amplification on blood obtained from a finger prick and found comparable allelic frequencies in somatic mutations between the finger prick sample and matched venous blood [101].

Aside from the clinical utility in humans, analysis of minute amounts of blood may facilitate longitudinal ctDNA monitoring from model organisms, such as rodents. Current methods for longitudinal monitoring are limited to high copy-number targets only, such as human LINE-1 sequences [102]. Potentially, sWGS of a blood spot obtained from a xenograft model followed by separate alignment to the rodent and human genome could permit detection of ctDNA and provide an alternative route to longitudinal monitoring from animal models.

Generating a cfDNA sequencing library from a blood spot is challenging due to the few cfDNA copies and abundance of long contaminating genomic DNA (gDNA) fragments present in the sample. A new method needs to be devised that removes the contaminating gDNA fragments while retaining the majority of the cfDNA fragments in the sample. An additional concern will be the low levels of DNA present in the sample and it remains to be seen if libraries can be prepared successfully from such limited inputs.

1.6 Thesis aims

The application of cfDNA for detection of cancer has been widely described in the literature and different methods have been developed. While these methods are great in detecting high levels of ctDNA in patients with a higher disease burden, they start to lose sensitivity when applied to samples with low levels of ctDNA and disease burden. Further research has indicated a potential to improve sensitivity by utilising methodological advances (e.g. UMIs) and biological features of cfDNA (e.g. fragment length). While some methods have since evolved that are utilising these features, they have not led to the desired increase in sensitivity.

Another hindrance in cfDNA analysis is the sample processing prior to analysis. The majority of the currently available methods require a stringent sample processing, with both skilled workers and the right equipment. Therefore, samples are mostly collected in large clinical centres that fulfil these requirements. If sample collection and processing were simplified, additional research centres and even general practices could participate in future studies, increasing the number of recruitable patients.

The work presented in this thesis intends to address both of these current limitations through:

1. The development of a tool for ctDNA detection and monitoring that utilises both technical advances and biological features to improve overall sensitivity.

2. The application of this tool to an independent cohort to assess the generalisability of the new method.

3. The development of a simplified method for collection and storage of blood samples, allowing less well equipped centres to participate in research studies.

Chapter 2

Patient-specific ctDNA monitoring from sequencing data

2.1 Attribution

This chapter is adapted from a manuscript, which was submitted to Science Translational Medicine in June 2019 and posted on bioRxiv in September 2019:
"High-sensitivity monitoring of ctDNA by patient-specific sequencing panels and integration of variant reads"
Jonathan C. M. Wan*, **Katrin Heider***,..., Charles Massie†, Pippa G. Corrie†, Nitzan Rosenfeld†,§

* Equally contributing authors
† CM, PGC and NR jointly supervised this work
§ Corresponding author

Together with Cancer Research Technologies Ltd we filed a patent based on the INVAR algorithm. Please see Chapter 6 for details.
A software package for the INVAR algorithm will also be made available on bitbucket. Please see Chapter 6 for details.

2.1.1 Author contributions

J.C.M.W., K.H. and N.R. wrote the manuscript. J.C.M.W., K.H., S.M., A.R-V., P-Y.C., G.R.B., C.A. and C.G.S. generated data. J.C.M.W., K.H., E.F., J.M., F.Mo., D.C. and N.R. developed the INVAR pipeline. J.C.M.W., K.H., J.M., E.F., A.M., W.Q., F.Ma., J.M. and D.C. analysed data and performed statistical analysis. A.B.G. and F.A.G. performed imaging

analysis. E.B., G.Y., I.H., W.N.C. and D.G. coordinated studies and participated in design. C.P., D.G., A.D., U.M., P.G.C. and N.R. led the MelResist study. P.C.G. and M.M. led the AVAST-M study. C.G.S., C.M., P.G.C., N.R. supervised the project. All authors reviewed and approved the manuscript.

Wet lab work

The wet lab work presented in this chapter was mostly performed by Jonathan Wan. Jonathan was a PhD student (2015-2018) who was mostly working on melanoma. He received the mutation calls for the late stage melanoma cohort which were previously called for Varela et al. [103]. For the stage II-III melanoma cohort, Jonathan organised the extraction of tumour tissue and buffy coats. He then prepared sequencing libraries and exome captures for these samples for identification of tumour mutations and submitted them for sequencing. Upon identification of tumour mutation lists for each of the patients, Jonathan designed and ordered the custom capture sequencing panels. Additionally, Jonathan organised the extraction of plasma samples and prepared libraries for the samples. He captured the libraries with the custom capture baits and submitted them for sequencing.

For the application of INVAR to whole exome sequencing, Jonathan and I jointly identified suitable cases (with high enough levels of ctDNA based on custom capture data) and split the workload for preparation and capture of the 21 samples. INVAR was also applied to shallow whole genome sequencing data. For these samples (n=33) Jonathan identified a subset of patients and time points from the larger cohort used for custom capture sequencing (above) and pooled remaining input libraries from the previous set of experiments for sequencing.

Dry lab work

The first iteration of the INVAR pipeline (thereafter v1) was developed by Jonathan Wan in our lab. Using v1 as a starting point, I developed the second iteration of the INVAR pipeline (hereafter v2) in close collaboration with Jonathan. As the scale of the project was such that we incorporated analysis of multiple cancer (melanoma, reported in this chapter, and lung cancer, reported in chapter 3) and data types (custom capture, whole exome and shallow whole genome sequencing), it was infeasible to conduct the pipeline development and application by myself. I would like to emphasise that the INVAR project was (at least) a two-person job as it required (often in parallel) to carry out additional experiments in the lab, update and / or trouble shoot the pipeline, generate and update the figures and write the paper. Hence, Jonathan and I are co-first authors on the publication presented in this chapter.

I attempted to quantify our contribution to the INVAR pipeline by summarising the commits per user to the tapas_pipeline Bitbucket repository which is shown in Fig. 2.1:

```
[C02NK1NYG3QQ:tapas_pipeline heider01$ git shortlog -s -n
   926  jonathancmw
   365  Katrin Heider
    32  Eyal Fisher
     7  James Morris
```

Fig. 2.1 **Screenshot of the users working on the v2 pipeline and their commits to the Bitbucket repository.**
The initial commits through Jonathan in this repository were made when the v1 pipeline was copied across to the new repository. Eyal Fisher was a mathematics MPhil student who joined our lab for a 3 months rotation during the pipeline development (Spring 2018) and was mostly responsible for the development of the underlying statistics. James Morris was a senior bioinformatician who mentored us throughout the pipeline development and helped in trouble shooting the code.

After data processing through the INVAR pipeline, Jonathan and I explored the output data further. This repository (tapas_paper) was mostly used for data visualisation and figure generation for the manuscript. In Fig. 2.2 I show the individual commits to this repository.

```
[C02NK1NYG3QQ:tapas_paper heider01$ git shortlog -s -n
   345  jonathancmw
   168  Katrin Heider
```

Fig. 2.2 **Screenshot of the users working on the data exploration and visualisation and their commits to the Bitbucket repository.**

Writing of the manuscript

Jonathan and I jointly devised a general structure for the manuscript and alternated writing the paper draft. We also jointly devised the figures to be presented for this work and took turns in plotting and adjusting them. After an initial draft we communicated with Nitzan to improve the appearance of both the text body and its figures. All final decisions regarding figures and text in the manuscript presented in this chapter were discussed with Jonathan and approved by Nitzan. I used this improved paper draft and incorporated it in this thesis, altering and repositioning figures and text where needed.

2.1.2 Competing interests

N.R. and D.G. are co-founders, shareholders and officers or consultants of Inivata Ltd, a cancer genomics company that commercialises ctDNA analysis. Inivata had no role in

the conceptualisation, study design, data collection and analysis, decision to publish or preparation of the manuscript. Cancer Research UK has filed patent applications protecting methods described in this manuscript.

2.1.3 Acknowledgments

The authors would like to thank Catherine Thorbinson, Alex Azevedo, Neera Maroo, from the MelResist and AVAST-M study groups, and the Cambridge Cancer Trial Centre, Addenbrookes Hospital.

2.1.4 Funding

We would like to acknowledge the support of The University of Cambridge, and Cancer Research UK (grant numbers A11906, A20240, and C2195/A8466). The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.337905.

2.2 Aims

The primary objective of this chapter was to develop a tool for ctDNA detection. Together with Jonathan Wan we investigated:

1. Which filtering steps should be applied to reduce the background error rate in (personalised) sequencing data. We explored the use of UMIs and the application of locus as well as sample specific filters to reduce the sequencing noise in a given sample.

2. Whether we can compute more accurate background error rates by analysing not only the observed mutation class but also its trinucleotide context.

3. Whether signal can be enhanced further by utilising sequencing information on cfDNA fragment length and allelic fraction information from the original tumour sample.

2.3 One sentence summary

Integrating tumour-derived sequences across large panels of patient-specific mutations offers enhanced sensitivity for ctDNA detection and monitoring from both high-depth and low-depth plasma sequencing data.

2.4 Abstract

Circulating tumour-derived DNA (ctDNA) can be used to monitor cancer dynamics noninvasively. Patients with small tumours have few copies of ctDNA in plasma, resulting in limited sensitivity to detect low-volume or residual disease. We show that sampling limitations can be overcome and sensitivity for ctDNA detection can be improved by massively parallel sequencing when hundreds to thousands of mutations are identified by tumour genotyping. We describe the INtegration of VAriant Reads (INVAR) analysis pipeline, which combines patient-specific mutation lists with both custom error-suppression methods and signal enrichment based on biological features of ctDNA. In this framework, the sensitivity can be estimated independently for each sample based on the number of informative reads, which is the product of the number of mutations analysed and the average depth of unique sequencing reads. We applied INVAR to deep sequencing data generated by custom hybrid-capture panels, and showed that when $\sim 10^6$ informative reads were obtained INVAR allowed detection of tumour-derived DNA fractions to parts per million (ppm). In serial samples from patients with advanced melanoma on treatment, we detected ctDNA when imaging confirmed tumour

volume of $\sim$1cm^3. In patients with resected early-stage melanoma, ctDNA was detected in 40% of patients who later relapsed, with higher rates of detection when more informative reads were obtained. We further demonstrated that INVAR can be generalised and allows improved detection of ctDNA from whole-exome and low-depth whole-genome sequencing data.

2.5 Introduction

Circulating tumour DNA (ctDNA) can be robustly detected in plasma when multiple copies of mutant DNA are present; however, when ctDNA levels are low, analysis of individual mutant loci might produce a negative result due to sampling noise even when using an assay with perfect analytical sensitivity. Such "missed" samples can have low fractional concentrations of ctDNA (relatively few mutant molecules in a high background), or low absolute numbers of mutant molecules due to limited sample input (Fig. 2.3). This effect of limited sampling reduces the sensitivity of ctDNA monitoring for patients with early-stage cancers, or following treatment for detection of minimal residual disease [8, 30]. Studies showed, for example, that by targeting a single mutation per patient in the plasma of early-stage breast and colorectal cancer patients post-operatively, ctDNA was detected in approximately 50% of patients who later relapsed [104, 105]. When applied to BRAF- or NRAS-mutant stage II-III patients with melanoma, ctDNA was detected up to 12 weeks post-surgery in only 16.8% of patients who relapsed within 5 years [106]. To increase the number of mutant molecules sampled, previous studies have shown that it may be possible to analyse larger volumes of plasma from multiple blood tubes [98, 105] and/or utilise broader sequencing panels.

Tumour-guided patient-specific analysis, which involves prior tumour genotype information and custom panel design [19, 35, 107, 108, 65], offers the possibility to greatly increase the sensitivity of ctDNA assays for cancer monitoring by targeting a larger number of mutations [8, 109] (Fig. 2.4). Such assays have analysed up to 40 patient-specific mutations in parallel, quantifying ctDNA to 1 mutant molecule per 25,000 copies in a patient with non-small cell lung cancer (NSCLC) [65]. Increasingly broad tumour sequencing is being performed both in research and clinical settings [110], which provides valuable mutation information that may be leveraged for improved sensitivity for ctDNA. We conceptualise the factors influencing ctDNA sensitivity as a two-dimensional space (Fig. 2.5A), highlighting the importance of maximising the number of relevant DNA fragments analysed, by increasing either plasma volumes or the number of (patient-specific) mutations sampled: the number of informative reads (IR) generated is proportional to the product of these two factors.

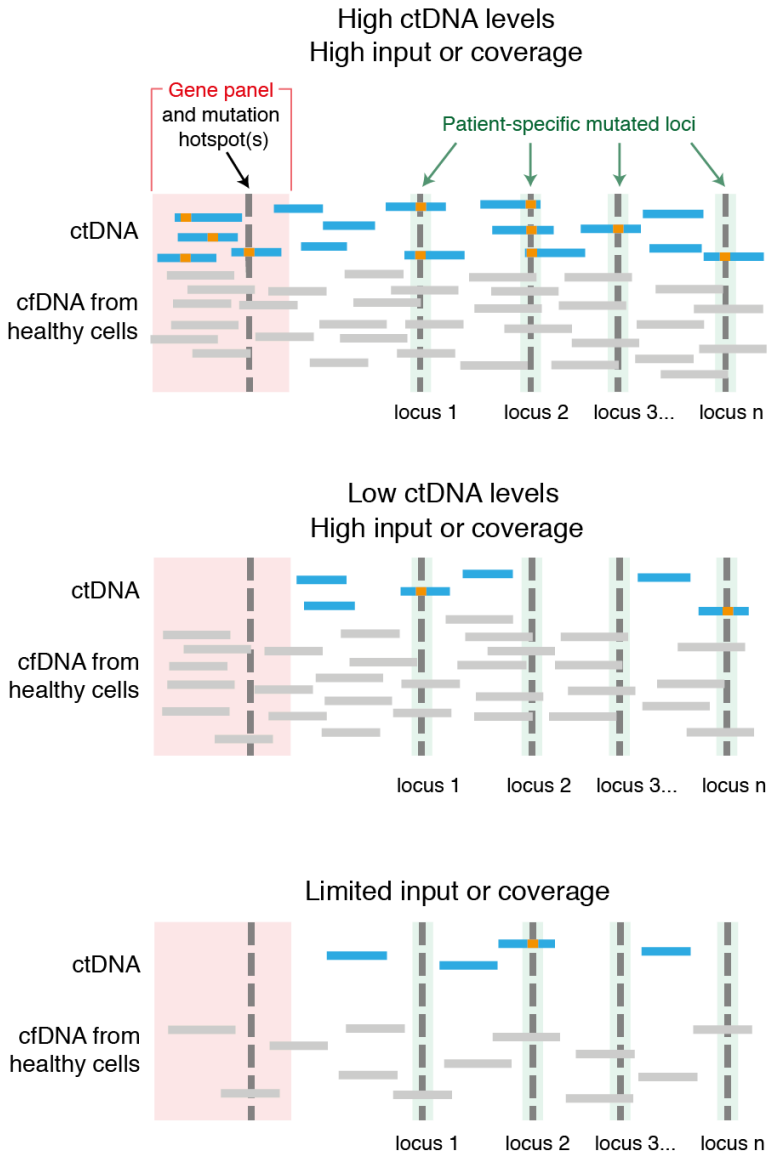


Fig. 2.3 **Patient-specific analysis overcomes sampling error in conventional and limited input scenarios.**
When high levels of ctDNA are present, gene panels and hotspot analysis are sufficient to detect ctDNA (top panel). However, if ctDNA concentrations are low these assays are at high risk of false negative results due to sampling noise. Utilising a large list of patient specific mutations allows sampling of mutant reads at multiple loci, enabling detection of ctDNA when there are few mutant reads due to either ultra-low ctDNA levels (middle panel), or due to limited starting material or sequencing coverage (bottom panel).

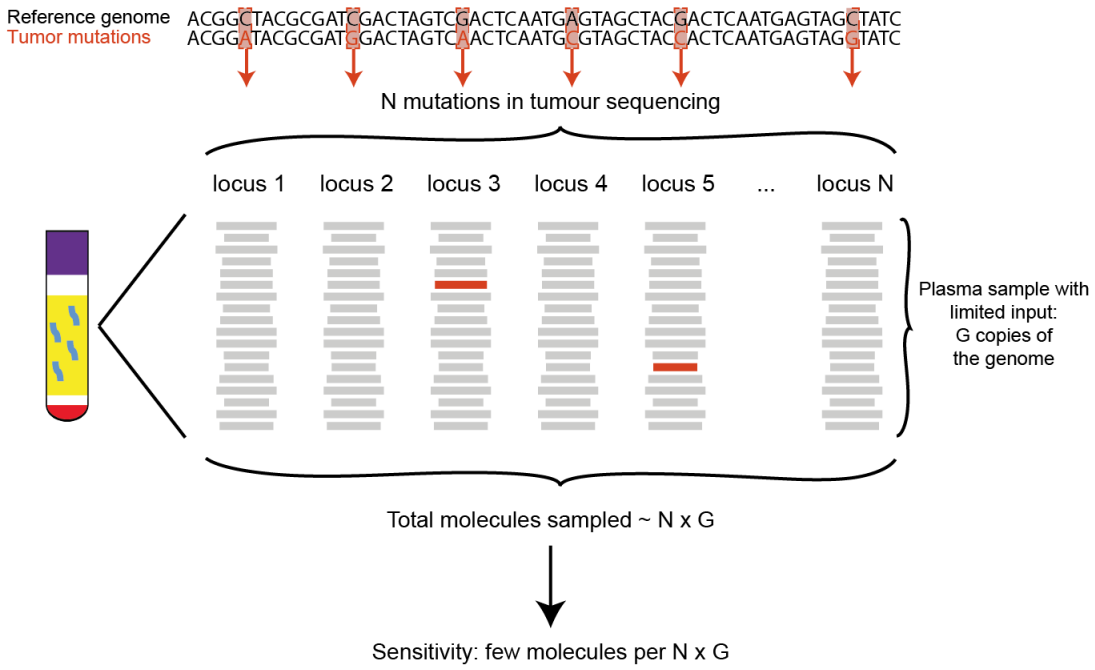


Fig. 2.4 **Targeting multiple mutations increases assay sensitivity.**
A given sample contains a limited number of copies of the genome, denoted by G. For plasma samples, the small amount of material limits the sensitivity that is attainable to one mutant copy in G total copies. By analysing in parallel a large number of marker loci (e.g. loci that are found to be mutated in the patient's tumour), denoted by N, detection of tumour DNA can be substantially enhanced to detect one or few mutant molecules per N x G copies. The same approach can be employed for other applications which aim to detect non-background/altered DNA, such as detection of foetal DNA or DNA from transplanted organs, in limited amounts of material such as plasma samples or other body fluids.

ctDNA detection methods often rely on identification of individual mutations [19, 64, 98] which may discard mutant signal that does not pass a threshold for calling. In this study, to improve sensitivity, we aggregated sequencing reads across 10^2-10^4 mutated loci, using prior information from tumour genotyping to guide analysis (Fig. 2.5B). The potential sensitivity benefit of targeting hundreds to thousands of tumour markers per patient has been previously suggested [65, 111], though such approaches have not been applied to cancer monitoring in plasma.

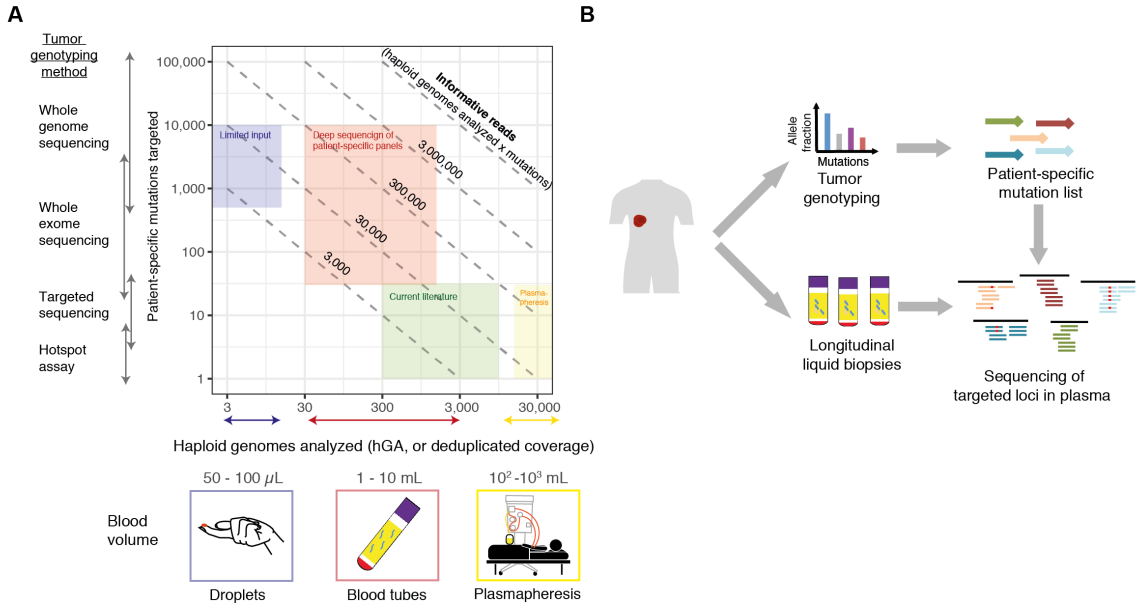


Fig. 2.5 **Study outline and rationale for integration of variant reads.**
(**A**)Illustration of the range of possible working points for ctDNA analysis using INVAR, plotting the haploid genomes analysed vs. the number of mutations. Diagonal lines indicate multiple ways to generate the same number of informative reads (IR, equivalent to haploid genomes analysed (hGA) x targeted loci). Current methods often focus on analysis of ∼10 ng of DNA (300-10,000 haploid copies of the genome) across 1 to 30 mutations per patient. This typically results in ∼10,000 IR, leading to frequently encountered detection limits of 0.01%-0.1% [19, 64]. mL, millilitre; µL, microlitre. (**B**) Patient-specific mutation lists generated by tumour genotyping were used to design hybrid-capture panels, that were applied to DNA extracted from plasma samples and yielded high depth sequencing. In later sections, the tumour genotyping data is used to analyse sequencing data from standard WES panels and shallow WGS.

We suggest that a tumour-guided approach targeting a large number of patient-specific mutations has advantages beyond simply mitigating sampling error. By virtue of generating a large number of IR, multiple error-suppression steps may be employed to overcome sequencing and PCR errors while retaining signal. Aside from molecular barcoding, it may

be possible to identify artefactual signal at a given locus by comparison of the given allelic fraction against the allele fractions at other patient-specific loci. Furthermore, greater weight may be assigned to fragments more likely to arise from tumour cells based on their biological characteristics such as fragment size [27], thereby enhancing the signal to noise ratio.

Here, we present a workflow for enhanced patient-specific monitoring that is optimised for sensitive detection of ctDNA using custom hybrid-capture panels (Fig. 2.6, flowchart in Fig. 2.7). This approach leverages custom error-suppression and signal enrichment methods to enable sensitive monitoring and identification of residual disease. We further demonstrate the ability to apply INVAR to plasma whole-exome sequencing (WES) and shallow whole genome sequencing (sWGS), demonstrating improved sensitivity for detection and quantification of ctDNA.

2.6 Tumour genotyping

First, tumour genotyping was performed to identify multiple patient-specific mutations per patient: exome sequencing data was generated from tumour and buffy coat samples from 47 patients with Stage II-IV melanoma (section 2.16), identifying a median of 625 mutations per patient (IQR 411-1076, Fig. 2.8 and Table 2.1). Mutations are shown by mutation class and trinucleotide context as well as observed tumour allele fraction (Fig. 2.9, Fig. 2.10). Interestingly, even after applying the filters suggested by Costello et al. [112], there are remaining acrtefactual mutations in the early stage melanoma cohort as seen by the relatively high number of C/A mutations, especially in the CCC and CCA contexts (Fig. 2.9). Additional filtering should be applied in the future to further remove these artefactual mutations.

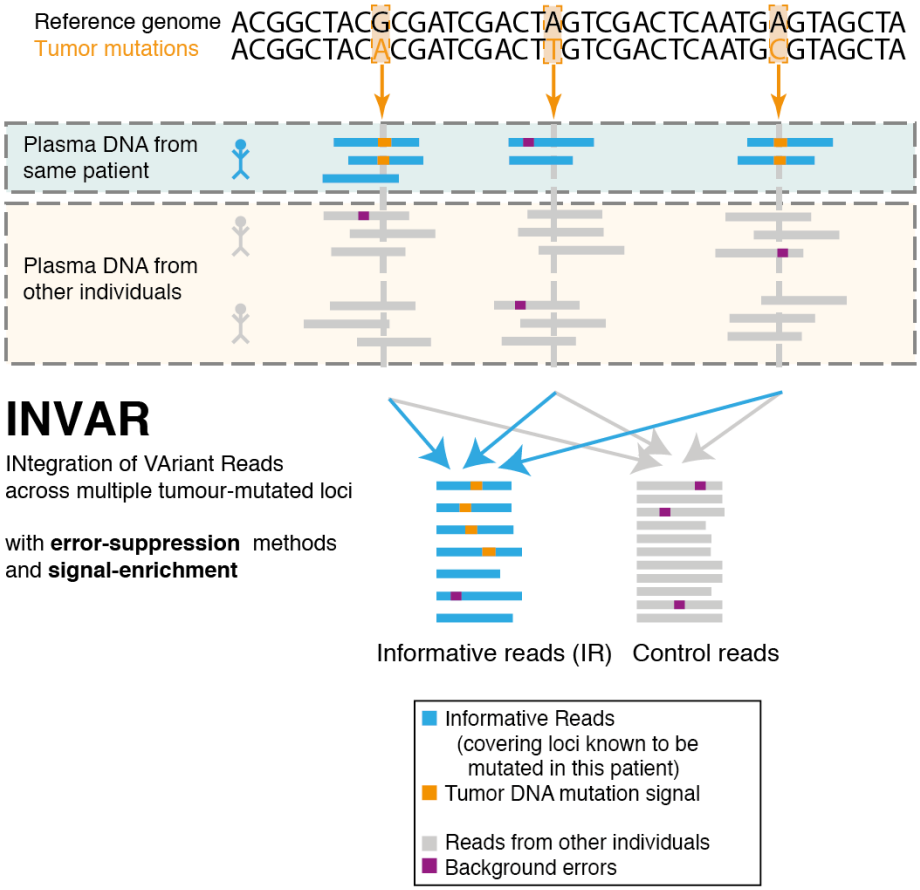


Fig. 2.6 **Overview of he INtegration of VAriant Reads (INVAR) pipeline.** To overcome sampling error, signal was aggregated across hundreds to thousands of mutations. Here we classify samples (rather than individual mutations) as significantly containing ctDNA, or not detected. 'Informative Reads' (IR, shown in blue) are reads generated from a patient's sample that overlap loci in the same patient's mutation list. Some of these reads may carry the mutation variants in the loci of interest (shown in orange). Reads from plasma samples of other patients at the same loci ('non-patient-specific') are used as control data to calculate the rates of background error rates (shown in purple) that can occur due to sequencing errors, PCR artefacts, or biological background signal. INVAR incorporates additional sequencing information on fragment length and tumour allelic fraction to enhance detection.

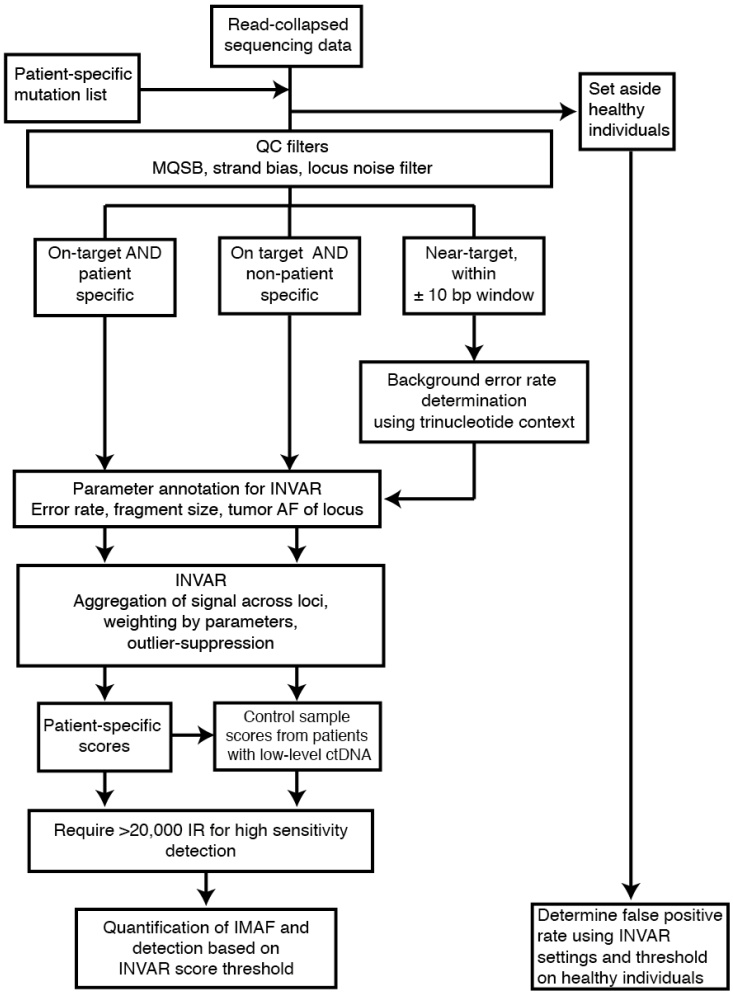


Fig. 2.7 **Flowchart of analysis steps in the INVAR pipeline.**
Integration of variant reads workflow. INVAR utilises plasma sequencing data and requires a list of patient-specific mutations, which may be derived from tumour or plasma sequencing. Filters are applied to sequencing data, then the data is split into: patient-specific (locus belonging to that patient), non-patient-specific (locus not belonging to that patient), and near-target (bases within 10 bp of all patient-specific loci). Patient-specific and non-patient-specific data are annotated with features that influence the probability of observing a real mutation. Outlier-suppression is applied to identify mutant signal inconsistent with the overall level of patient-specific signal. Next, signal is aggregated across all loci, considering annotated features, to generate an INVAR score per sample. Based on non-patient-specific samples, an INVAR score threshold is determined using ROC analysis for each cohort. Healthy control samples separately undergo the same steps to establish a specificity value for each cohort.

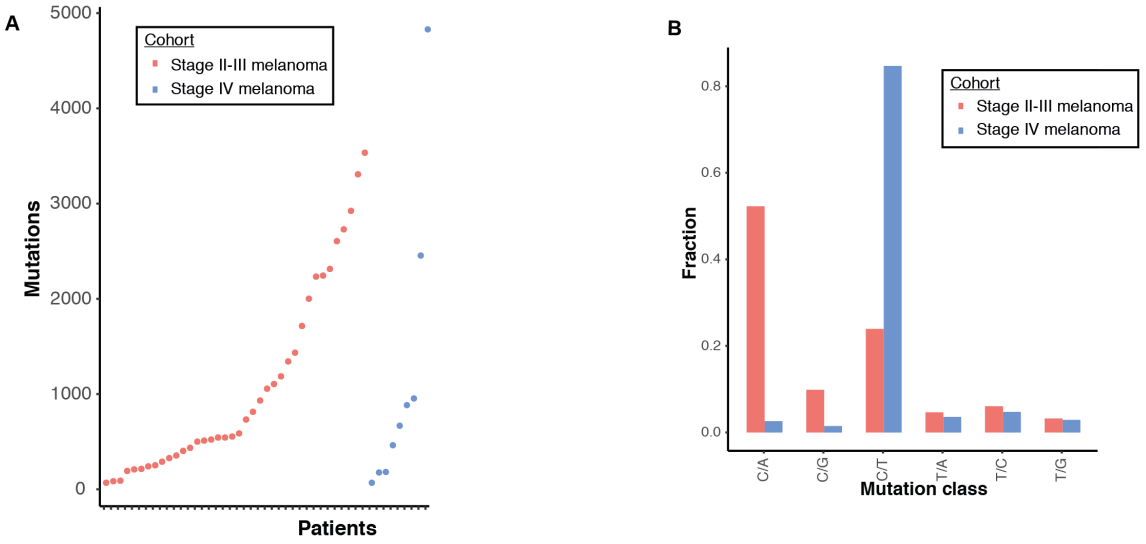


Fig. 2.8 **Tumour mutation list characterisation for INVAR.**
(**A**) Number of somatic mutations per patient, ordered by cohort. (**B**) Frequency of each mutation class included in each panel design.

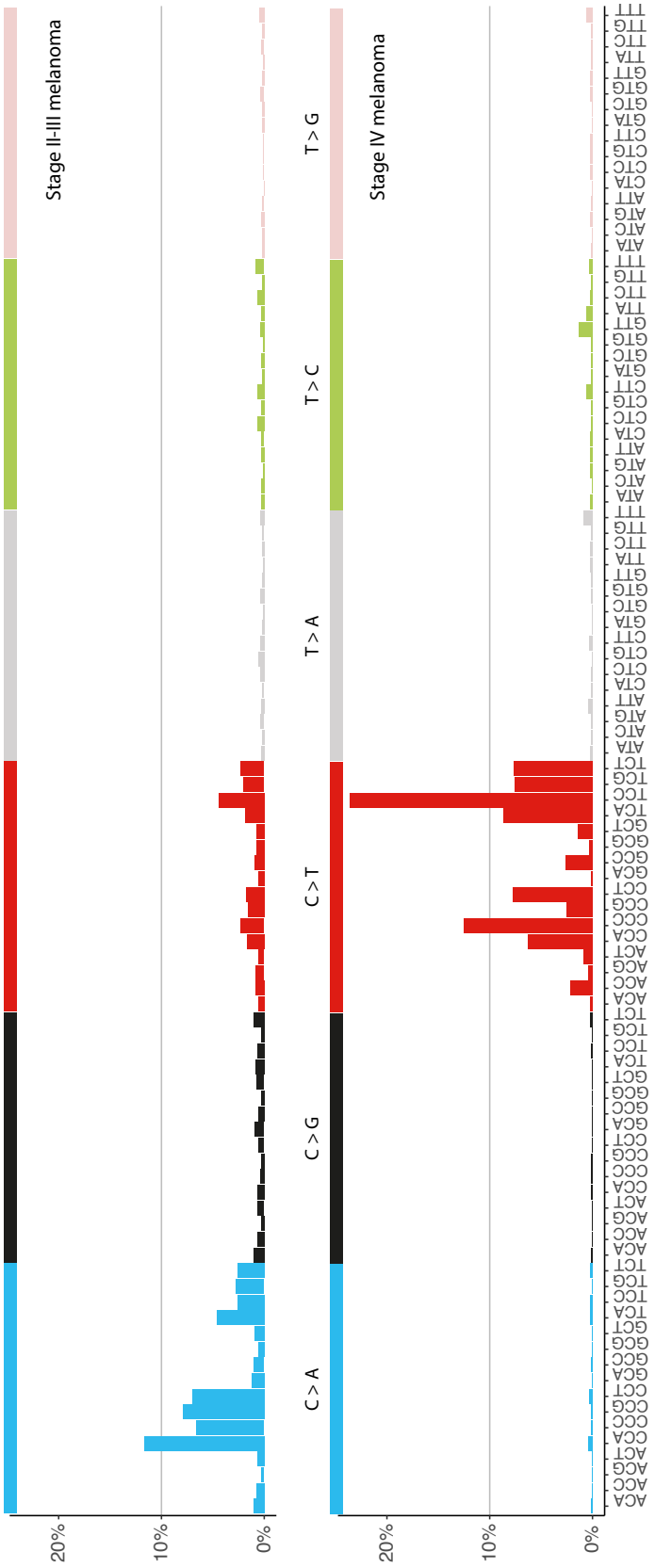


Fig. 2.9 **Trinucleotide context for tumour mutations.**
Proportion of mutations for each trinucleotide context and mutation class, grey line indicates 10% of all mutations in the cohort.

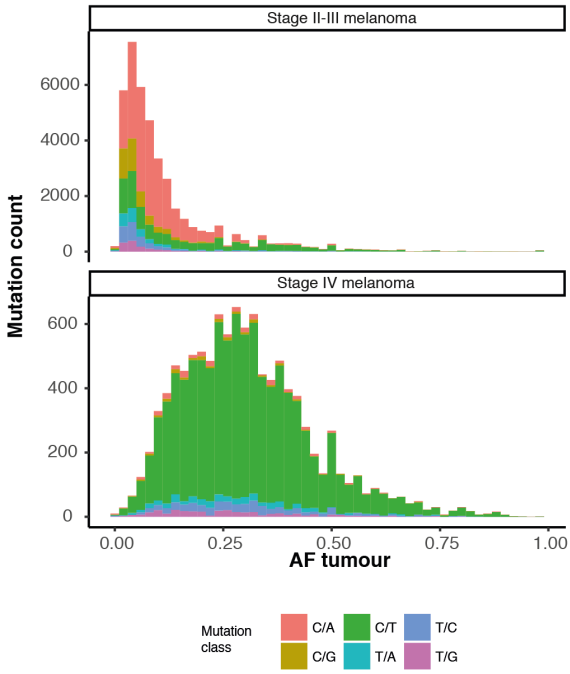


Fig. 2.10 **Distribution of tumour mutation allele fractions per cohort coloured by mutation class.**

We analysed the abundance of the commonly mutated genes in melanoma patients in our cohort. Based on the literature, BRAF (51%), NRAS (28%), NF1 (14%), CDKN2A (13%) and TP53 (15%) are the most commonly mutated genes in melanoma [113]. In the present cohort we see a higher representation of BRAF mutations (61%, Fig. 2.11). This is unsurprising as patients in this cohort are receiving BRAF-targeted therapy. Inerestingly, we also see a higher abundance of NF1 mutations (32%). However, as described by Akbani and colleagues, NF1 mutations are correlated to BRAF non-hot-spot mutations. We see mutations in both BRAF and NF1 in 11 of 19 patients with NF1 mutations, likely explaining the higher representation of NF1 mutations in this cohort (Fig. 2.11) [113]. The observed mutation rates of CDKN2A (18%), TP54 (15%) and NRAS (15%) are overall comparable to those seen in the literature (Fig. 2.11).

2.7 Characterising background error rates

We started by characterising background error rates in hybrid-capture sequencing data. Approximation of error rates may potentially be achieved through grouping mutations of similar class. We demonstrate that error rates vary between mutation class by over an order of magnitude using raw sequencing data without using molecular barcodes (Fig. 2.12A),

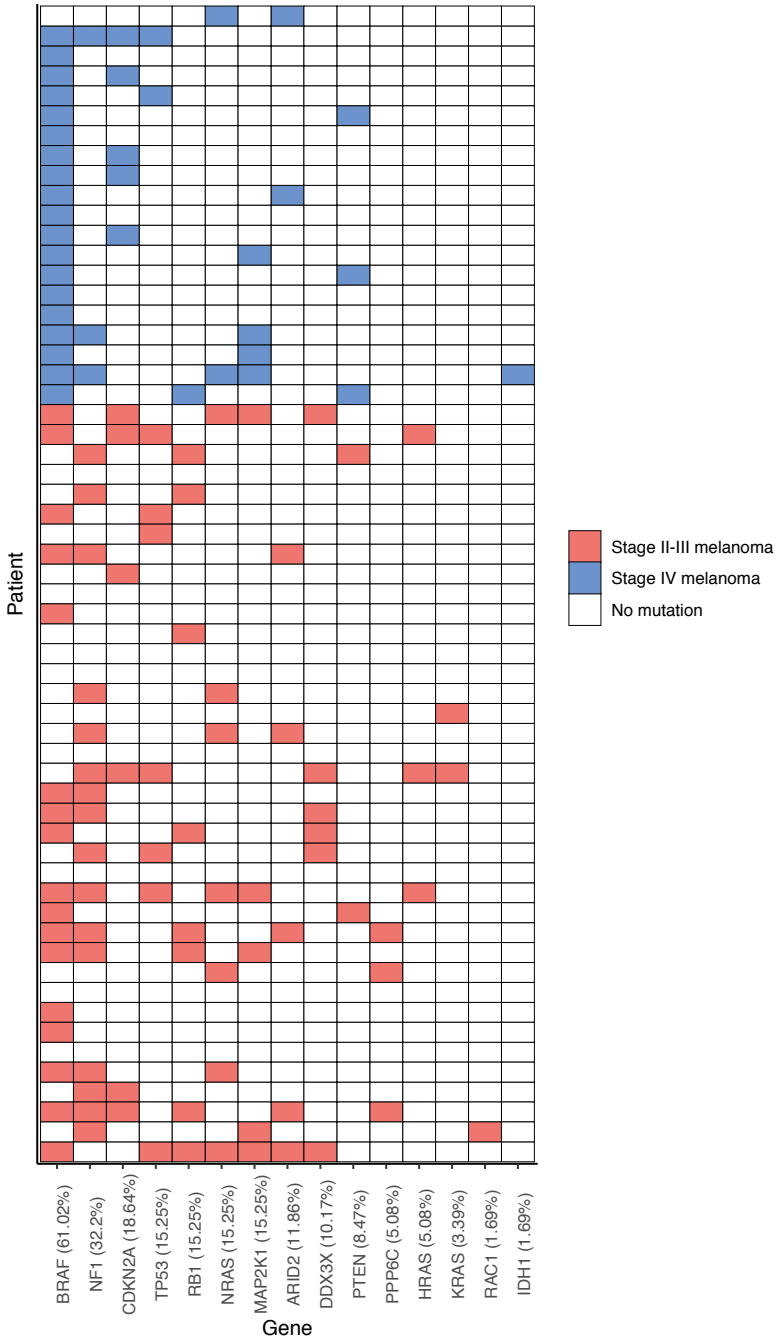


Fig. 2.11 **Mutated genes in melanoma cohort.**Most commonly mutated melanoma genes were identified from [113]. Boxes are coloured if patients had a mutation in a given gene. Blue fields indicate a patient belonging to the stage IV melanoma cohort while red boxes indicate a patient belonging to the stage II-III melanoma cohort. White boxes indicate that no mutation was present.

consistent with Newman et al. [65]. To increase the resolution of background error rates further, we grouped mutations by both mutation class and trinucleotide context, demonstrating over two orders of magnitude difference in background error rate between the least and most noisy trinucleotide context (Fig. 2.12B).

Using a patient-specific sequencing approach, a large number of private mutation loci were targeted. Each locus has its own error rate, though accurate benchmarking of the background noise rates of individual loci to levels below 10^{-6} would require the cfDNA molecules from a total of 100mL of plasma in order to sample one mutant read. This assumes a cfDNA concentration of 10ng/mL from plasma, yielding 3 million analysable molecules. Thus, we sought to develop a background error model for patient-specific sequencing data that could estimate the background error rate of a locus accurately using limited control samples. In this study, 99.8% of the mutations identified by tumour sequencing were private i.e. unique to each individual. We assessed if patients may be used to control for other patients' mutation lists, thereby enabling us to group patient-specific mutation lists of multiple patients together and reduce the number of additional control samples to be run on each panel. In this study, a mean of 5.5 patients were included on each custom hybrid-capture sequencing panel design. There was no significant difference in background error rate whether using healthy individuals or other patients serving as controls ('patient-control' samples, which may control for other patients at private loci) (Fig. 2.12C). Thus, INVAR utilises sequencing data from one patient to control for others with both custom and untargeted approaches such as exome or WGS (Fig. 2.12D).

2.8 Error-suppression in patient-specific sequencing data

As part of the INVAR pipeline, we sought to develop methods to minimise artefacts in patient-specific sequencing data. Read collapsing was performed using unique molecular barcodes which reduced error rates across all mutation classes (Fig. 2.13A), similar to previous studies [58]. Increasing the minimum number of duplicates required per read family reduced error rates further, but at the expense of a greater fraction of the sequencing data being discarded (Fig. 2.13B). To balance data loss against background error rate, a minimum family size threshold of 2 was used.

INVAR requires any mutation signal to be represented in both the forward (F) and reverse (R) read of the read pair. This serves to both reduce sequencing error and produces a small size-selection effect for short fragments since only short fragments would be read completely in both F and R with paired-end 150bp sequencing. This step retained 92.4% of mutant reads and 84.0% of wild-type reads in a training dataset (Fig. 2.13C).

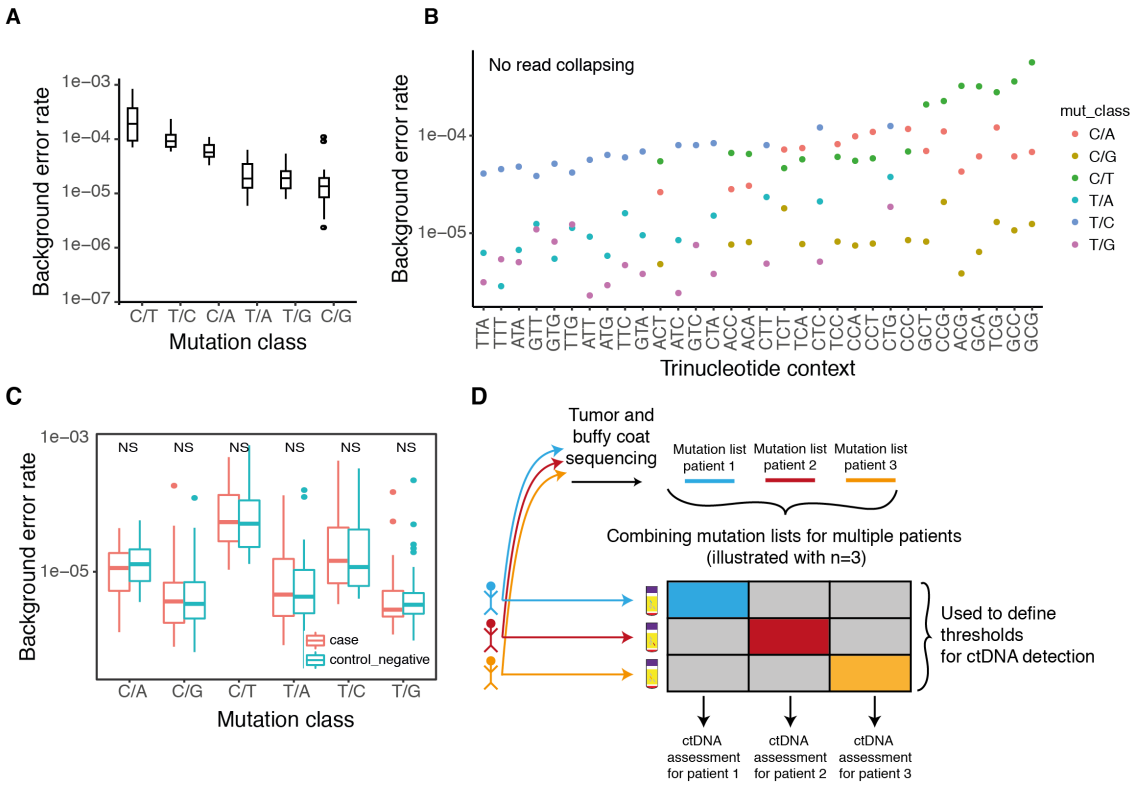


Fig. 2.12 **Characterisation of background error rates.**
(**A**) Non-error suppressed background error rates by mutation class. Trinucleotide error rates are grouped by mutation class. Data from 136 samples were used for this analysis. (**B**) Non-error suppressed background error rates by trinucleotide context. (**C**) Background error rates were calculated by mutation class for healthy control individuals (blue) and patient samples (red) after equalising the number of read families per group. Complementary mutation classes were combined. T-tests were performed between healthy and patient samples. Error rates obtained from patient samples are not significantly different from the error rates in healthy control samples. NS, not significant. (**D**) Overview of the usage of sequencing data by the INVAR pipeline. Data is collected for each locus of interest in the matched patient (shown by coloured boxes), and control data is obtained by analysing the same loci in additional patients from the same cohort for whom the loci were not found to be mutated in the tumour or buffy coat analysis. Such data can be generated by applying a standardised sequencing approach, such as WES/sWGS, to all samples (Fig. 2.28) or by combining multiple patient-specific mutation lists into a custom capture panel that is sequenced across multiple patients (Fig. 2.21). Data from other patients in those loci (‘non-matched mutations’) are used to determine background mutation rates and a ctDNA detection cut-off.

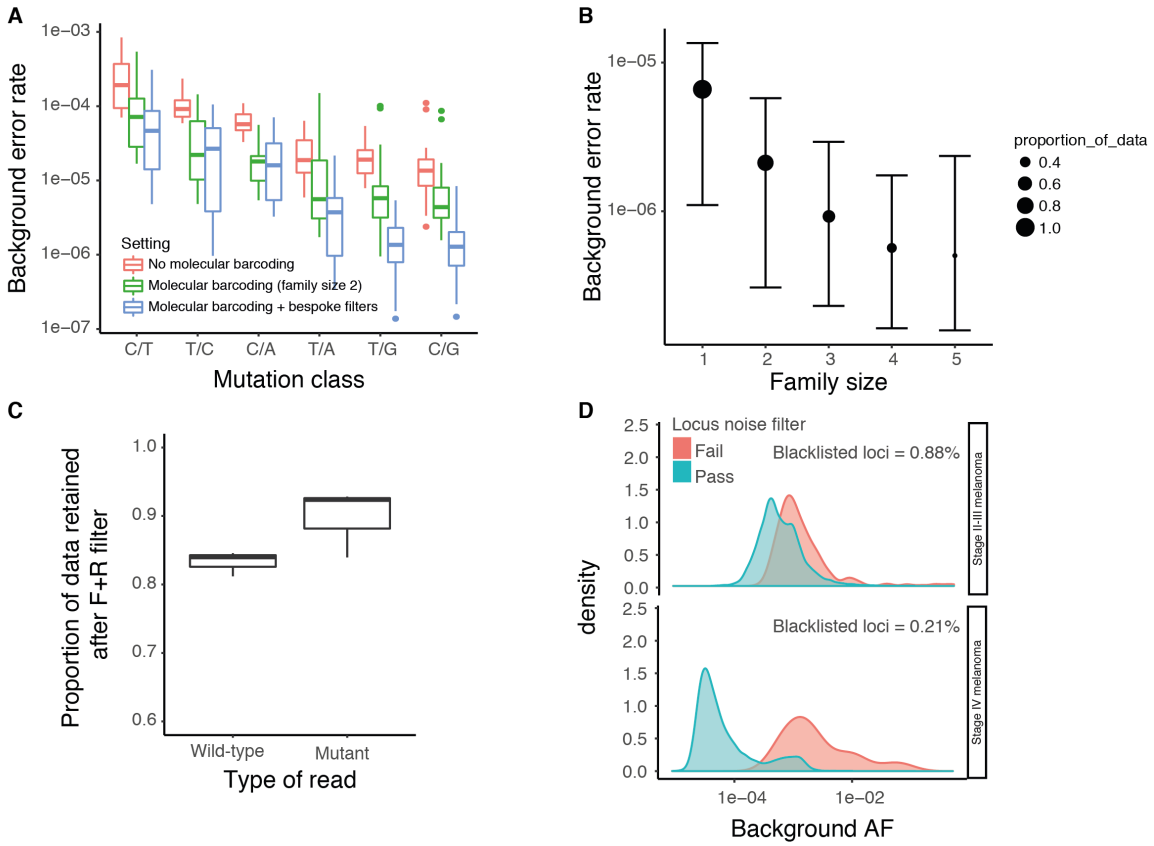


Fig. 2.13 **Effect of read collapsing and locus noise filter on background error.** (**A**) Comparison of background error rates for non-error suppressed (red) and error suppressed (green, minimum family size of 2) data. Error rates of error suppressed data can be further improved through bespoke INVAR filters (blue). Background error rates were calculated by aggregating all non-reference bases across all considered bases. (**B**) Overall background error rates resulting from different minimum family size requirements, and the proportion of read families retained with each setting. (**C**) Effect of requiring forward and reverse reads at a locus; a median of 84.0% of the wild-type reads and a median of 92.4% of the mutant reads were retained with this filter. (**D**) Background error rates were characterised per locus based on all reads generated from control samples, split by cohort. Loci passing the locus noise filter are shown in blue, those not passing the filter are shown in red. The proportions of loci blacklisted by this filter are indicated at the top right.

When targeting a large number of patient-specific loci, it becomes increasingly likely that technically noisy sites, or single-nucleotide polymorphism (SNP) loci are included in the list. Newman et al. have previously utilised position specific background polishing to address this issue [65]. In this study we blacklisted loci that showed either error-suppressed mutant signal in >10% of the patient-control samples, or a mean background error rate of >1% mutant allele fraction. This approach excluded 0.5% of the patient-specific loci (Fig. 2.13D). Requiring mutant signal in both reads and applying a locus noise filter reduced noise modestly when applied individually; however, when combined they showed a synergistic effect, reducing background error rates to below 10^{-6} in some mutation classes (Fig. 2.15A and Fig. 2.14A). The individual effects of these filters on individual trinucleotide contexts are shown in Fig. 2.14B.

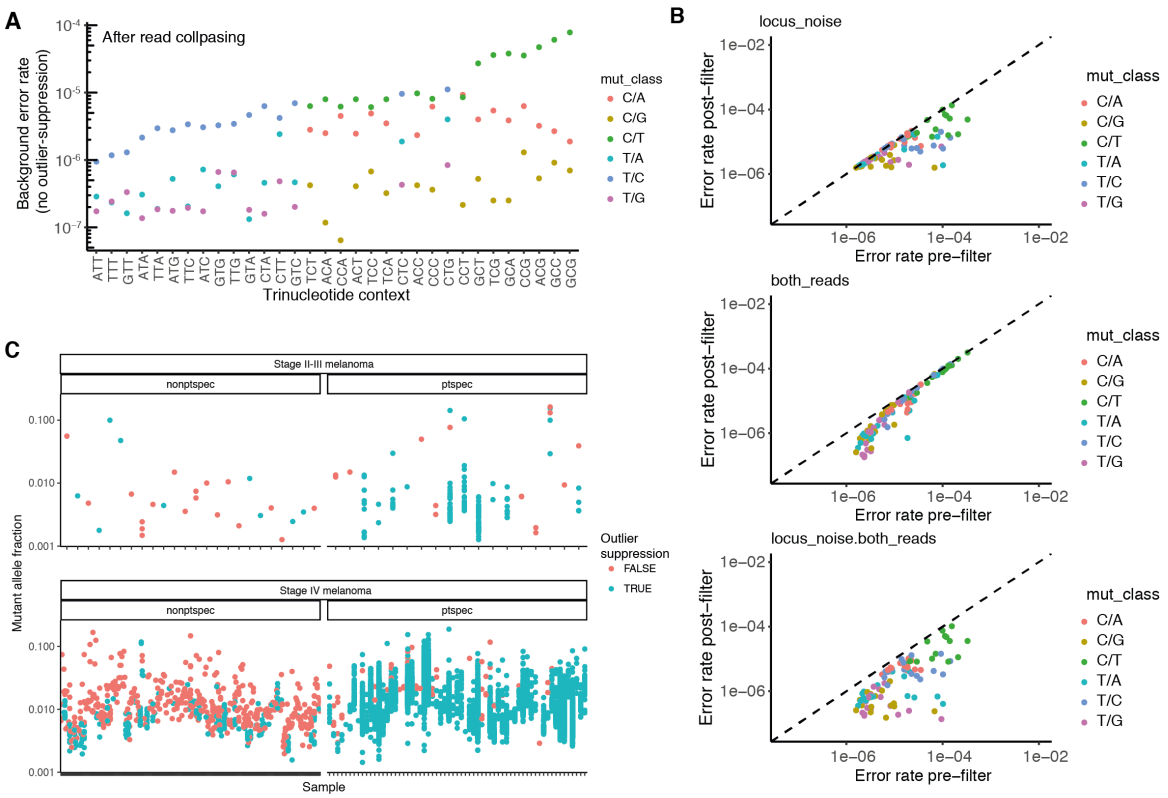


Fig. 2.14 **Analysis of trinucleotide error rates and patient outlier suppression.** (**A**) Error rates by trinucleotide context and mutation class following data filtering but before applying the patient specific outlier suppression. (**B**) For each trinucleotide context, background error rates are plotted before and after each background error filter, highlighting the collective benefit of each of the error filters. (**C**) Outlier suppression filter: Raw data-points for both cohorts (showing patient and control samples), with the outlier-suppressed data points indicated in red (Details in section 2.17).

When targeting a large number of patient-specific sites, it becomes possible to assess the distribution of allele fractions observed. In the residual disease setting, we expect to have a high degree of sampling error. Therefore, signal should appear stochastically as individual mutant molecules are distributed across patient-specific loci, with many of the loci having zero mutant reads. In order to optimise INVAR for detection of the lowest possible levels of ctDNA, we developed a method called patient-specific outlier suppression to exclude signal at a locus that is not consistent with the remaining loci (Fig. 2.14C and Fig. 2.15B). This tests each locus against the distribution of signal at all other loci with a correction for multiple testing, excluding loci that are significantly outlying. Mutant signal was reduced 3-fold in control samples, while retaining 96.1% of mutant signal in patient samples (Fig. 2.15C).

Overall, combining the above steps results in an average 131-fold decrease in background error relative to raw sequencing data (Fig. 2.15A) and reduces the error rates of some trinucleotide contexts to below $1x10^{-6}$ (Fig. 2.14A).

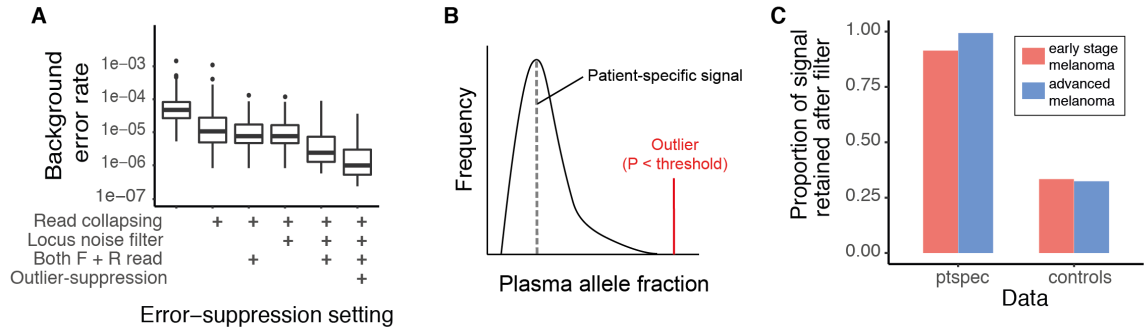


Fig. 2.15 **Development and analytical performance of the INVAR method.**
(**A**) Reduction of error rates following different error-suppression settings (section 2.17). (**B**) Loci observed with significantly greater signal than the remainder of the loci of that patient might be due to noise at that locus, contamination, or a mis-genotyped SNP locus (in red, see section 2.16). (**C**) Summary of effect of outlier suppression on both cohorts. Mutant signal was reduced 3-fold in control samples, while retaining 96.1% of mutant signal in patient samples.

2.9 Patient-specific signal enrichment

To enhance detection further, INVAR is able to enrich for ctDNA signal through probability weighting based on the tumour allele fraction of each mutation locus and ctDNA fragment sizes. Tumour mutations with a higher tumour allelic fraction are more likely to be observed in the plasma [108, 114], therefore, greater weight was allocated to mutant signals in plasma

from loci with high tumour mutant allele fraction. Using a dilution series, we confirm the relationship between the tumour allele fraction of a locus and rate of detection of ctDNA of that locus in plasma (Fig. 2.16A). We confirm in clinical samples that patient-specific mutation loci observed in plasma had a significantly higher tumour allele fraction compared to those not observed in plasma ($P = 2x10^{-16}$; Wilcoxon test, Fig. 2.17D).

Analysis of 144 samples showed a nucleosomal pattern of cfDNA fragmentation, with mutant fragments shorter than wild-type fragments at the mono-nucleosome and di-nucleosome peak (Fig. 2.16B). We also observed that stage IV melanoma patients had a significantly higher median mutant fragment size compared to the stage II-III melanoma patients (163bp vs. 154bp, $P = 2x10^{-16}$, Wilcoxon test, Fig. 2.16C). Previous research has shown enrichment for ctDNA when shorter fragments are selected using either in vitro or in silico size selection [14, 15, 27]. However, at low levels of signal, such methods can cause loss of rare mutant alleles [12]. Thus, in this study we weighted each signal based on its fragment size in order to boost ctDNA signal, while retaining all the data (Fig. 2.17E). Based on smoothed size profiles of mutant and wild-type fragments observed (Fig. 2.16D), patient data were used to size-weight other patients' data using a leave-one-out approach (section 2.17). Following signal weighting, INVAR aggregates signal across all patient-specific mutations (section 2.16). In order to determine whether or not ctDNA is detected in a sample, data from non-matched mutations in other patients were used as negative controls to set the detection threshold (Fig. 2.18). An Integrated Mutant Allele Fraction (IMAF) is determined by taking a background-subtracted, depth-weighted mean allele fraction across the patient-specific loci in each sample (section 2.17).

2.10 Analytical sensitivity and specificity of INVAR

To benchmark the sensitivity of INVAR, we performed custom capture sequencing of a dilution series of plasma from one melanoma patient (stage IV disease), for whom we identified 5,073 mutations through exome sequencing. Plasma DNA from this patient was serially diluted into control volunteers' plasma DNA to an expected IMAF of 3.6×10^{-7}. Without use of unique molecular barcodes, INVAR detected ctDNA down to an expected allele fraction of 3.6×10^{-5}, which was quantified to an average IMAF of 4.7×10^{-5} from two replicates (Fig. 2.19A). Following the use of molecular barcodes and custom error-suppression methods, the diluted ctDNA was detected to an expected IMAF of 3.6×10^{-6} (3.6 parts per million) in two replicates, with IMAF values of 4.3 and 5.2 ppm. Overall the correlation between IMAF and the expected mutant fraction was 0.98 (Pearson's r, $p < 2.2 \times 10^{-16}$, Fig. 2.18A). At an expected allele fraction of 3.6×10^{-7}, ctDNA was detected in 2

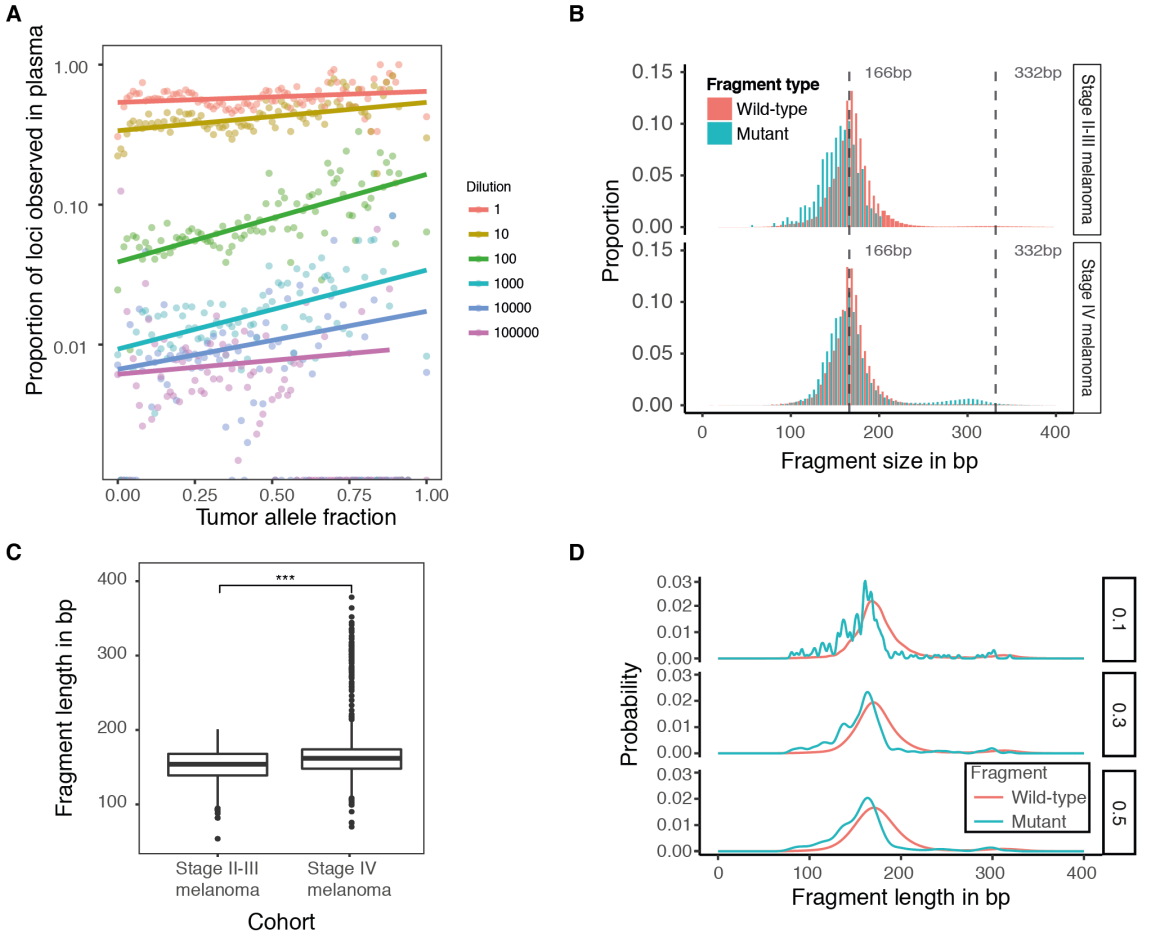


Fig. 2.16 **Utilising tumour allelic fraction information and plasma DNA fragment length to enhance ctDNA signal.**
(**A**) Comparison of tumour and plasma mutant allele fractions. Using error-suppressed data, tumour loci were grouped into bins of 0.01 mutant allele fraction, and the proportion of loci observed in plasma was determined for different levels of a dilution series. The dilution level of the spike-in dilution series is indicated by each colour. At each dilution level, there is a positive correlation between the tumour allele fraction and proportion of loci observed in plasma. (**B**) For each cohort, size profiles were generated for mutant and wild-type fragments. (**C**) Comparison of mutant fragment distributions between cohorts. These were compared using a two-sided Wilcoxon rank test after downsampling the number of mutant reads to match for both cohorts. (**D**) The distributions of fragment sizes for different levels of smoothing, used to assign weights to fragment sizes (section 2.17).

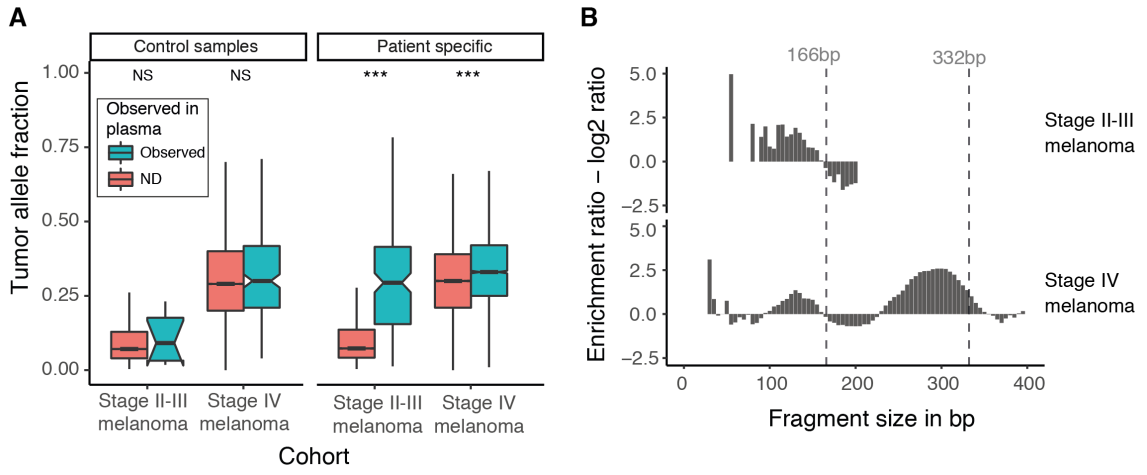


Fig. 2.17 **Patient-specific signal enrichment in INVAR.**
(**A**) Tumour allele fractions were compared between loci with and without detected signal in plasma. Loci with signal in plasma had significantly higher tumour allele fractions in patient samples. There was no significant increase in tumour allele fraction when performing this analysis on non-patient-specific samples (Student's t-test, NS, not significant; *** = P < 0.0001). (**B**) Log2 enrichment ratios for mutant fragments from three different cohorts of patients. Size ranges enriched for ctDNA are assigned more weight by the INVAR pipeline.

out of 3 replicates. To assess the impact on sensitivity of the number of mutations targeted we downsampled sequencing data in silico to include subsets of patient-specific mutation lists. This confirmed that targeting more mutations resulted in more IR and correspondingly higher ctDNA detection rates (Fig. 2.19B, section 2.17).

The false positive rate of INVAR was measured twice, once in patient-control samples and separately in healthy control samples. First, analytical specificity was determined through analysis of samples from other patients (patient-control samples) at non-matched mutation loci, giving a median specificity of 98.0% (Fig. 2.18, Table 2.4). To confirm the specificity of INVAR in independent control samples, we ran custom capture sequencing (with the same oligo pools) on samples from healthy individuals and analysed those by INVAR using each of the patient-specific mutation lists. The ROC curve for the stage IV melanoma cohort controlled against healthy individuals is shown in Fig. 2.19C. Across each of the analyses in this study, using control cfDNA from 26 healthy individuals, a median specificity value of 97.05% was obtained, consistent with the analytical specificity defined in non-matched control samples from other patients (Fig. 2.20).

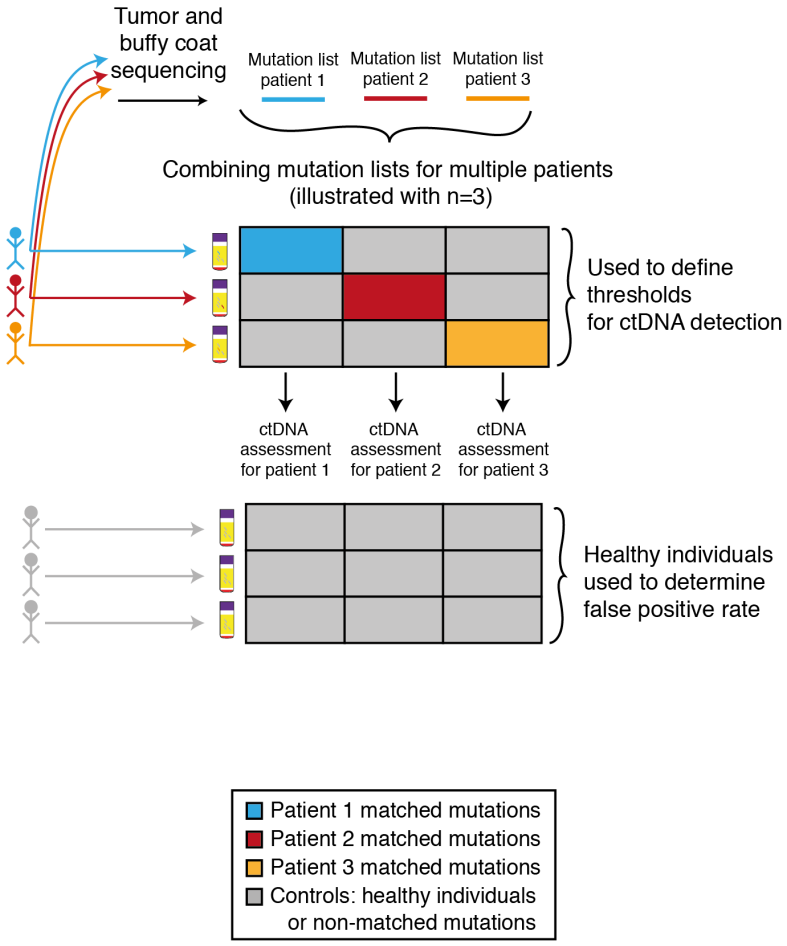


Fig. 2.18 **Overview of the INVAR pipeline.**
INVAR leverages patients to control for one another and uses separate healthy controls. In this study, individual mutation lists are generated from tumour and buffy coat sequencing. Each locus of interest is sequenced in the matched patient, and in additional patients from the same cohort for whom this locus was not found to be mutated in the tumour or buffy coat analysis. This can be done by applying a generic panel to all samples (such as WES/WGS, Fig. 2.28), or by combining multiple patient-specific mutation lists into a combined custom panel that is sequenced across multiple patients (Fig. 2.21). For each patient, INVAR aggregates the sequencing information across the loci of the patient-specific mutation list. Data from other patients in those loci ('non-matched mutations') are used to determine background mutation rates and detection cut-offs (section 2.17). Additional samples from healthy individuals are analysed by the same panels, this data was used to assess the false positive rates in healthy individuals.

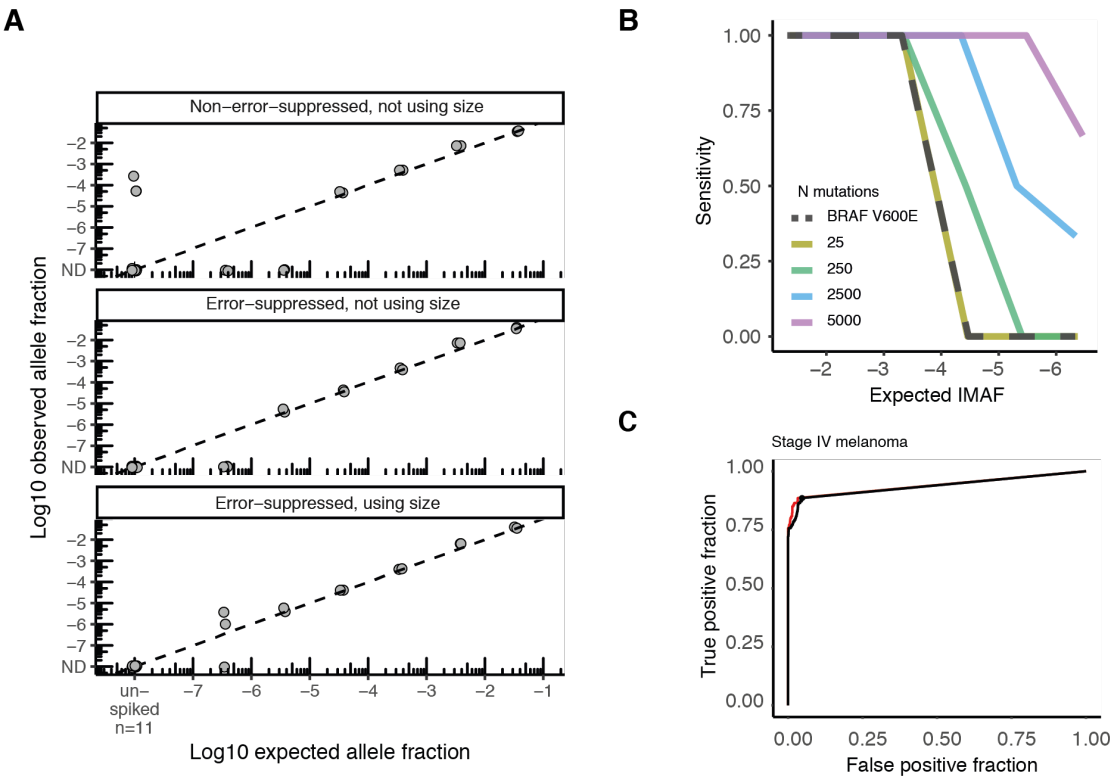


Fig. 2.19 **Sensitivity and specificity determination of INVAR.**
(**A**) Spike-in dilution experiment to assess the sensitivity of INVAR. Using error-suppressed data with INVAR, ctDNA was detected in replicates for all dilutions to 3.6 ppm, and in 2 of 3 replicates at an expected ctDNA allele fraction of 3.6 x 10^{-7} (section 2.17). Using error-suppressed data of 11 replicates from the same healthy individuals without spiked-in DNA from the cancer patient, no mutant reads were observed in an aggregated 6.3 x 10^{6} informative reads across the patient-specific mutation list. (**B**) The sensitivity in the spike-in dilution series was assessed after the number of analysed loci was downsampled in silico to between 1 and 5,000 mutations (section 2.17). (**C**) ROC analysis of the stage IV melanoma cohort was performed against patient-controls (black) and healthy individuals (red).

A

Cohort	Analytical spec.	Specificity
Stage IV melanoma	95.4%	96.7%
Stage II-III melanoma	98.4%	N/A
WES plasma	96.1%	95.8%
sWGS plasma	100%	97.6%

B

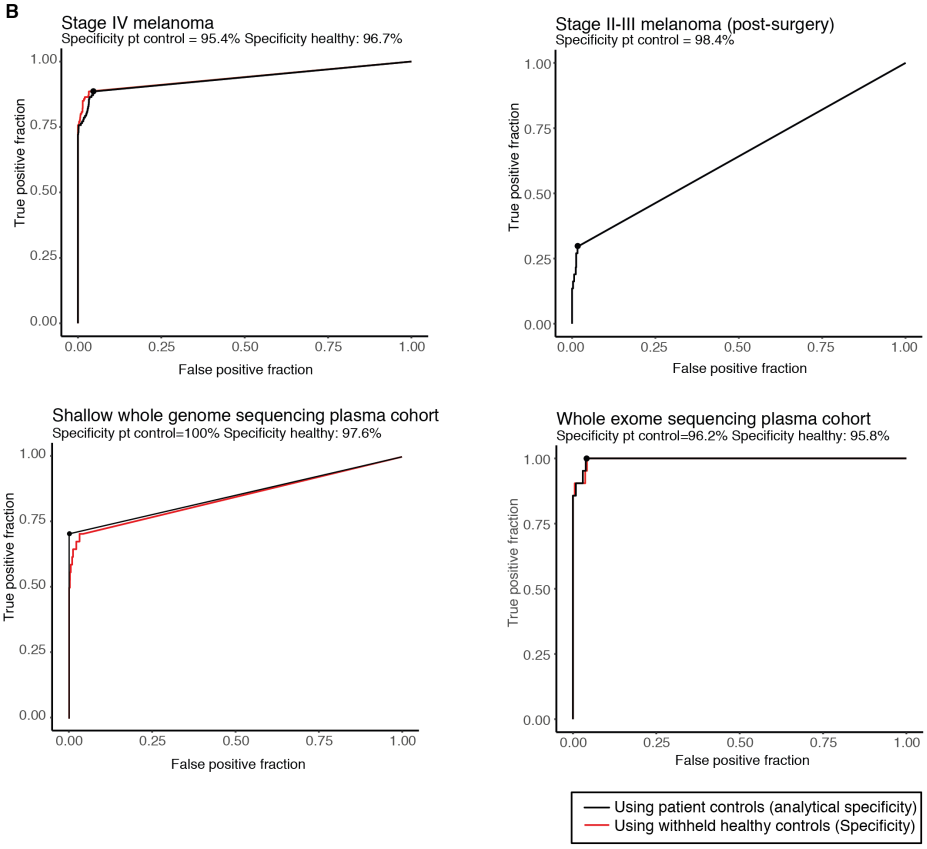


Fig. 2.20 **ROC curves and specificity for all cohorts and data types.**
(**A**) Analytical (based on patient control samples) and clinical specificity values (based on healthy individuals) are provded for all cohorts. No healthy individuals were tested on the custom capture panel of the stage II-III melanoma cohort. (**B**) For all cohorts, ROC analyses were performed against patient-controls (black) and healthy individuals (red). For the stage II-III melanoma (post-surgery) cohort, our analysis was blinded to outcome, and patients who did not relapse within 5 years were also included in the ROC analysis; thus, the maximal possible sensitivity for this cohort (as defined) was the fraction of relapsing patients (18/33 = 54.5%). INVAR detected 8 out of 20 patients who relapsed (ROC showing sensitivity at 9/33 = 27.3%).

2.11 Quantification of ctDNA in patient samples

We applied INVAR to custom capture panel sequencing data from 130 plasma samples from 47 stage II-IV melanoma patients, generating up to 2.9 x 10^6 IR per sample (median 1.7 x 10^5 IR), thus analysing orders of magnitude more cfDNA fragments compared to methods that analyse individual or few loci (Fig. 2.21A). In this study, we demonstrated a dynamic range of 5 orders of magnitude and detection of trace levels of ctDNA in plasma samples (Fig. 2.21B and C); this detection was obtained from a median input material of 1638 copies of the genome (5.46 ng of DNA; Table 2.2). In a total of 13 of the 130 plasma samples analysed with custom capture sequencing, ctDNA was detected with signal in fewer than 1% of the patient-specific loci (Fig. 2.21D). The lowest fraction of cancer genomes detected was 1/714, equivalent to <5 femtograms of tumour DNA. Given the limited input, the low ctDNA levels detected would be below the 95% limit of detection for a 'perfect' single-locus assay in 48% of the cases (indicated with filled circles in Fig. 2.21C). The input mass vs. IMAF of each sample is shown in Fig. 2.22A, highlighting the sensitivity benefit of a broad sequencing approach. Thus, targeting multiple mutations can allow detection of low absolute amounts of tumour-derived DNA.

In Stage IV melanoma patient samples, ctDNA IMAF values showed a correlation of 0.8 with tumour size assessed by CT imaging (Pearson's r, P = 6.7 x 10^{-10}, Fig. 2.22B, Table 2.5), comparable to other studies [19, 31]. Similarly, ctDNA IMAF had a correlation of 0.53 with serum lactate dehydrogenase (LDH), a routinely used clinical marker for monitoring of melanoma (Pearson's r, P = 2.8 x 10^{-4}, Fig. 2.23). INVAR analysis was used to monitor ctDNA dynamics in response to treatment, in which the majority of patients received anti-BRAF targeted therapy first line, which resulted in a rapid decline in ctDNA in those patients (Fig. 2.24). In one patient (#59) treated with a series of targeted therapies and immunotherapy, ctDNA was detected down to an IMAF of 2.5 ppm, corresponding to a radiological tumour volume of 1.3 cm^3 (Fig. 2.25). Following progression on vemurafenib, patient #59 progressed on multiple other anti-BRAF targeted therapies (pazopanib, dabrafenib and trametinib) and immunotherapy (ipilimumab), corresponding to a constant rise in ctDNA over two years of monitoring (Fig. 2.25).

The IMAF obtained from INVAR in the late stage melanoma cohort was compared to allelic fractions obtained from TAm-Seq [35] based amplicon sequencing of a single mutation locus (targeting BRAF or NRAS depending on the mutation status of the patient). The two methods correlated well down to allelic fractions of 1 x 10^{-2} (Pearson's r = 0.85, P = 1.9 x 10^{-15}, Fig. 2.26). However, the single mutation amplicon approach is unable to detect ctDNA below 1 x 10^{-2} while INVAR detects ctDNA down to 1 x 10^{-6}, highlighting the advantage of INVAR over a single locus assay (Fig. 2.26).

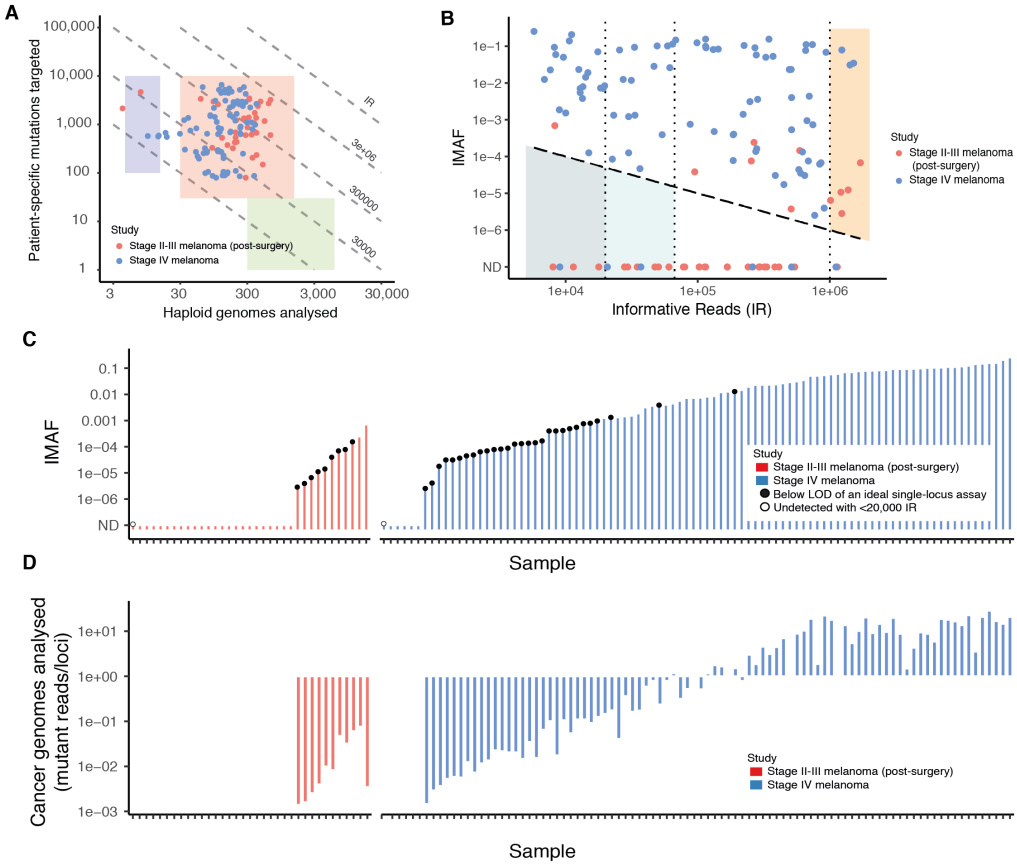


Fig. 2.21 **ctDNA detection by INVAR in early and advanced disease.**
(**A**) Number of haploid genomes analysed (hGA) and the number of mutations targeted in 130 plasma samples from 47 cancer patients across two cohorts. Dashed diagonal lines indicate the number of targeted loci x hGA that yield the indicated IR. Shaded regions highlight the working point of current methods (green), the working point of INVAR from custom capture (red) and sWGS plasma data (blue) (**B**) Two-dimensional representation of detected ctDNA fractions are plotted against the IR for each sample. Samples falling above the dashed line, which is plotted at 1/IR were detected. In some samples, $>10^6$ IR were obtained, and ctDNA was detected down to fractions of few ppm (orange shaded region). We used a sensitivity threshold of 20,000 IR (left-most dotted line), such that samples with undetected ctDNA with fewer than 20,000 IR (detection level 0.01%) were called "unclassified" and excluded from the analysis (total of 4 of 130 samples; dark blue shaded region). Using an alternative threshold, for example 66,666 IR (indicated by the second dotted line and the light blue shaded region), will increase the overall detection rates in the cohorts. (**C**) ctDNA fractional levels (IMAF) detected in this study, shown in ascending order for the two cohorts. Filled circles indicate samples where the number of hGA would fall below the 95% limit of detection for a perfect single-locus assay given the measured IMAF (section 2.17). Empty circles indicate unclassified samples. (**D**) Copies of the cancer genome detected for each of the samples (same order as above in part (C)), calculated as the number of mutant fragments divided by the number of loci queried (Table 2.2).

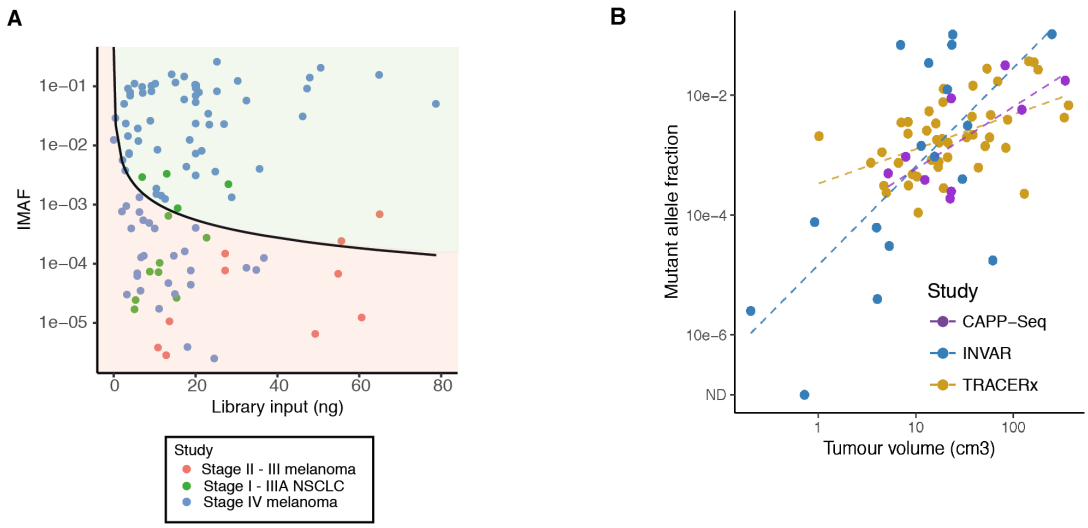


Fig. 2.22 **Characterisation of ctDNA levels in advanced melanoma.**
(**A**) Comparison of input mass and IMAF observed. For each library with detected ctDNA, the DNA input mass for library preparation is plotted against IMAF. The black line indicates the threshold below which a perfect single-locus assay would have <95% sensitivity. In this study, 48% of samples would not be detectable using a perfect single-locus assay with the plasma DNA input amounts used. (**B**) Comparison between ctDNA and tumour volume in our study (Pearson's r = 0.67, P = 0.003) and in previous publications measuring multiple mutations per patient in NSCLC, using CAPP-Seq (Pearson's r = 0.72, P = 0.003) [31], and multiplexed PCR in the TRACERx cohort (Pearson's r = 0.5, P = 4.3 x 10^{-4}) [19].

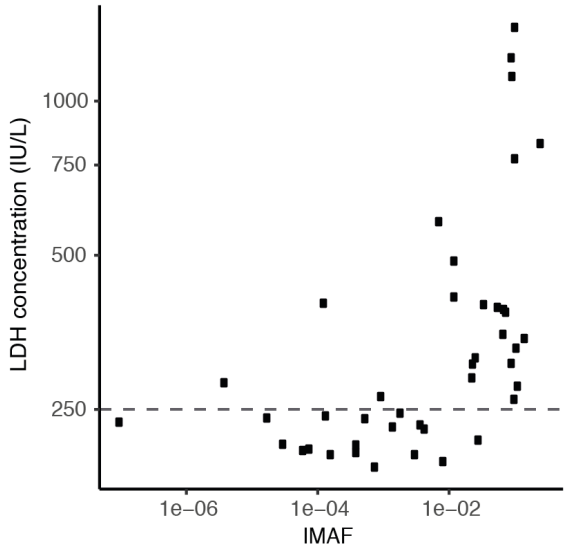


Fig. 2.23 **Relationship between serum lactate dehydrogenase and IMAF in advanced stage melanoma patients.**
Pearson's r = 0.46, P = 0.0058. The dashed line at 250IU/L indicates the upper limit of normal for LDH.

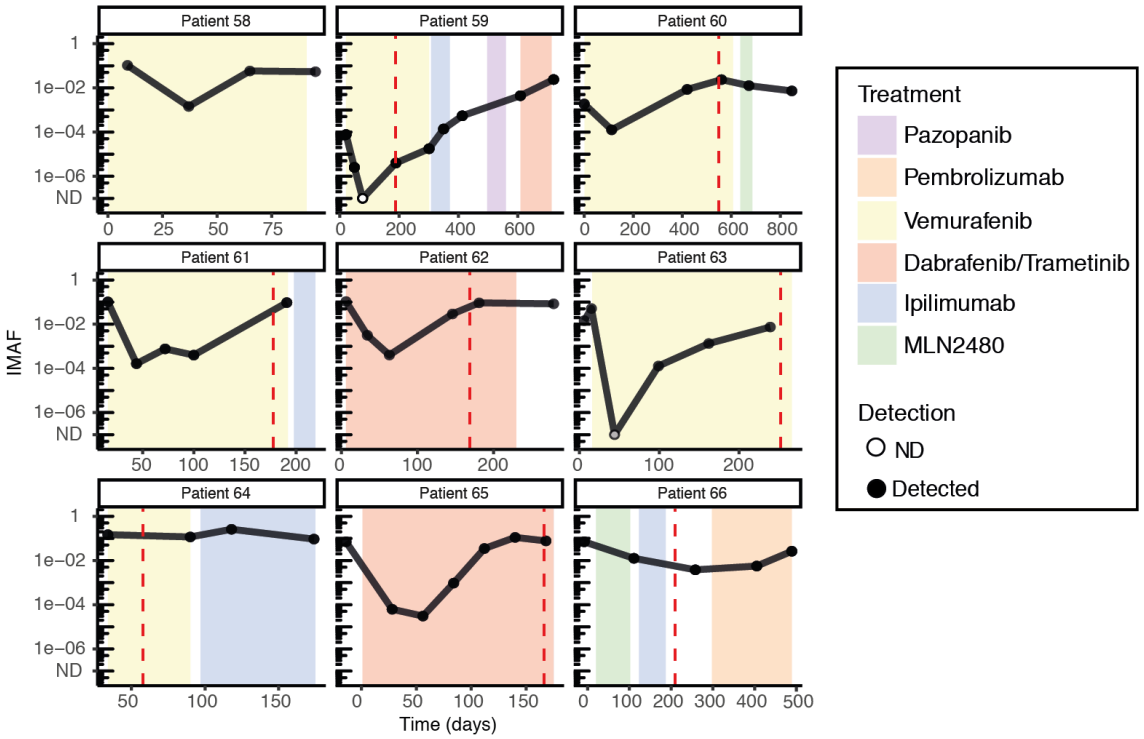


Fig. 2.24 **Longitudinal ctDNA profiles for advanced melanoma patients.**
IMAF values are plotted over time per patient, using error-suppressed individualised sequencing data. Vertical dashed lines indicate time of radiological progression.

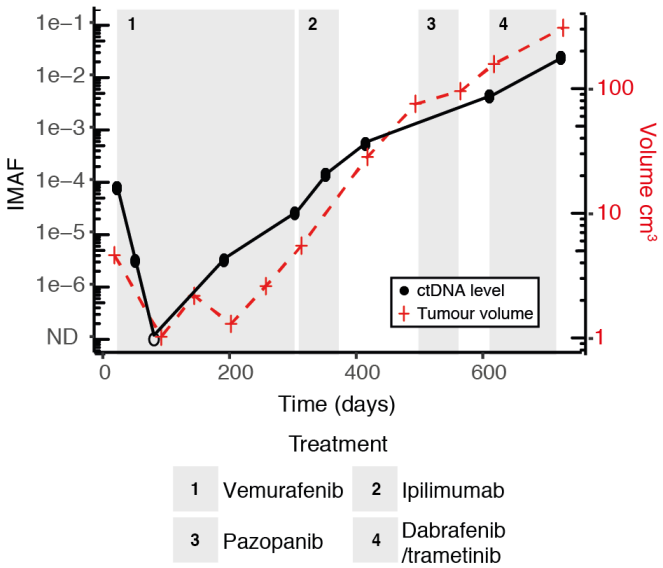


Fig. 2.25 **Longitudinal monitoring in late stage melanoma patient.**
ctDNA IMAF and tumour volume plotted over time for one patient with metastatic melanoma over the course of several treatment lines (indicated by shaded boxes). ctDNA was detected to 2.5 ppm during treatment with anti-BRAF targeted therapy, when disease volume was approximately 1.3 cm^3.

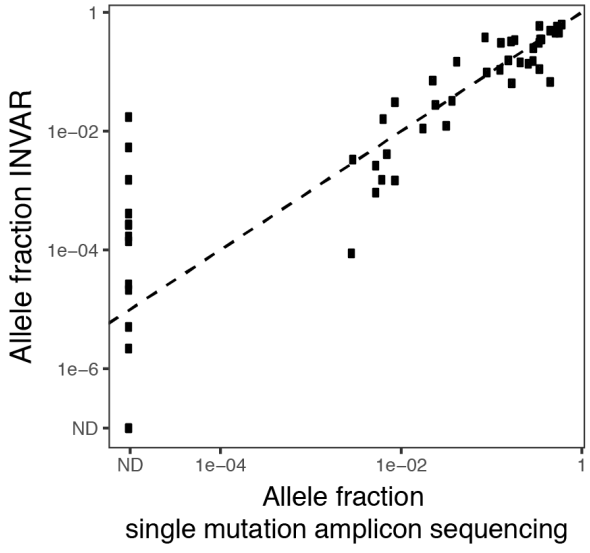


Fig. 2.26 **Comparison of INVAR and single locus assay.**
IMAF obtained from INVAR was compared to the allelic fraction obtained from a single locus TAm-Seq assay, showing good correlation down to 1 x 10^{-2} (Pearson's r = 0.85, P = 1.9 x 10^{-15}). INVAR continues to detect ctDNA down to 1 x 10^{-6}, while these samples remain undetected with the single locus assay. ND, not detected.

2.12 ctDNA detection post-surgery

To test INVAR in the residual disease setting, we applied INVAR to post-operative samples from 38 patients with resected Stage II-III melanoma recruited in the UK AVAST-M trial. Patient samples were collected up to 6 months after surgery with curative intent. The clinical details of this cohort are given in Fig. 2.27A. We interrogated a median of 3.6 x 10^5 IR (IQR 0.64 x 10^5 to 4.03 x 10^5) and detected ctDNA to a minimum IMAF of 2.85 ppm, indicated in Fig. 2.21C. The specificity of this analysis was >0.98 (Fig. 2.20). In total, ctDNA was detected in samples from 11 of 38 patients (28.9%). ctDNA was detected at higher rates when higher numbers of informative reads were obtained, with ctDNA detected in 10 of 28 (35.7%) cases with >66,666 informative reads (sensitivity of detection of 30 ppm), 9 of 18 (50%) cases with >250,000 informative reads (sensitivity of detection below 10 ppm) and in 5 of 6 (83%) cases with >10^6 informative reads (Fig. 2.21B). Samples with no ctDNA detected and few informative reads may indicate limited resolution and would benefit from additional information (more informative reads, obtainable from deeper sequencing or more mutations). A similar approach was previously described in the relative haplotype dosage method by Lo et al. [9]. In our case, excluding 3 samples where ctDNA was undetected and had fewer than 20,000 informative reads (sensitivity of detection of 0.01% not reached), ctDNA was detected in 8 of 20 (40%) patients who later recurred and was associated with a strong trend for shorter disease-free interval (6.3 months vs. median not reached with 5 years' follow-up; Hazard ratio (HR) = 2.08; 95% CI 0.85-5.13, P = 0.11) and overall survival (2.6 years vs. median not reached, P = 0.08). In comparison, a previous analysis of ctDNA detection at 12 weeks after surgery in 161 patients with resected BRAF or NRAS mutant melanoma detected ctDNA in 16.8% of patients who later relapsed [106].

2.13 Sensitive ctDNA monitoring using WES and sWGS

Patient-specific capture panels allow highly sensitive detection of ctDNA, but require prior design of patient specific capture panels. Therefore, we assessed whether INVAR could be applied to standardised workflows such as WES or WGS. This allows the panel design step to be omitted and requires only the patient-specific mutation list from tumour sequencing, which may be performed in parallel with plasma sequencing to save time (Fig. 2.28A).

To test the generalisability of INVAR, we selected samples with IMAF values quantified as being between 4.5 x 10^{-5} and 0.16 using custom-capture sequencing and utilised commercially available exome capture kits to sequence plasma DNA to a median raw depth of 238x. Despite the modest depth of sequencing, we obtained between 1,565 and 473,300 IR

A

	Total	
	N	%
Characteristics		
	38	
Sex		
Male	17	45
Female	21	55
Breslow thickness		
<=2.0mm	16	42
>2-4.0mm	8	21
>4.0mm	12	32
Unknown	2	5
Ulceration		
Present	9	24
Absent	23	60
Unknown	6	16
Disease stage		
II	8	21
IIIA	4	10
IIIB	17	45
IIIC	9	24
N classification		
II (No or N/A)	8	21
III (N1a and N2a)	6	16
III (other N)	24	63
ECOG performance status		
0	35	92
1	3	8
BRAF and NRAS mutation		
BRAF mutation	19	50
NRAS mutation	8	21
BRAF and NRAS WT	10	26
BRAF WT and NRAS not tested	1	3
Trial arm		
Observation	38	100
Time from latest surgery to trial entry in weeks		
Median (Range)	11 (4-12)	

B

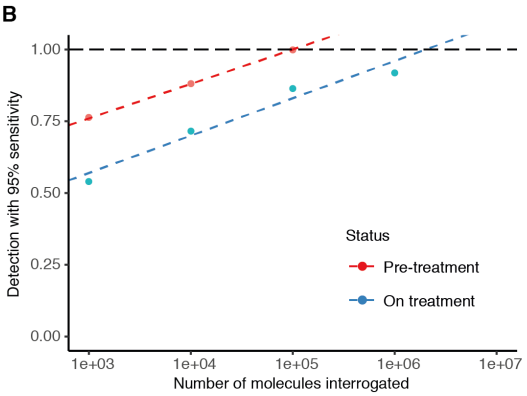


Fig. 2.27 **Characterisation of IMAF values in the early-stage melanoma cohort.** (**A**) Summary table of patient characteristics for the stage II-III resected melanoma cohort (n = 38). (**B**) We estimated the detection rates of ctDNA for different levels of IR (section 2.17). We observe a linear relationship ($R^2 = 0.95$) between the number of IR and detection rate in the baseline samples of the stage IV melanoma cohort. ctDNA was detected in 100% of baseline samples with 10^5 IR (red), whereas following the initiation of treatment, 10^6-10^7 IR are needed to detect all longitudinal samples (blue), reflecting the lower levels of ctDNA.

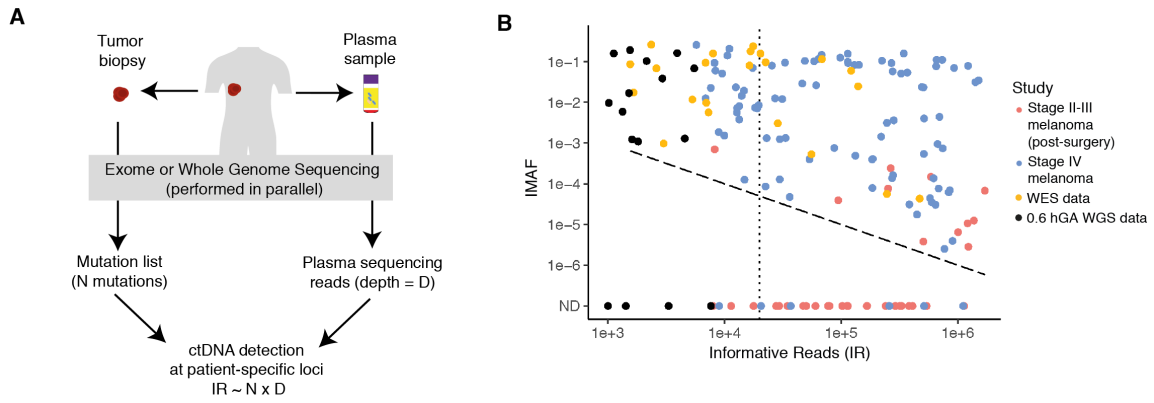


Fig. 2.28 **Application of INVAR to WES/WGS data.**
(**A**) schematic overview of a generalised INVAR approach. Tumour (and buffy coat) and plasma samples are sequenced in parallel using whole exome or genome sequencing, and INVAR can be applied to the plasma WES/WGS data using mutation lists inferred from the tumour sequencing. (**B**) INVAR was applied to WES data from 21 plasma samples (238x raw depth), and to WGS data from 33 plasma samples (0.6x raw depth). For each sample IMAF values are plotted against the number of unique IR and compared to the application of custom capture sequencing. Both sWGS and WES applications of INVAR can detect ctDNA but show a lower number of IR compared to custom capture data. Dotted vertical line: 20,000 IR threshold, dashed diagonal line: 1/IR.

using WES (Fig. 2.28B). We detected ctDNA in all 20 samples tested down to IMAFs as low as 4.34 x 10^{-5} (Fig. 2.30A), demonstrating that ctDNA can be sensitively detected by INVAR from WES data using patient-specific mutation lists. These IMAF values showed a correlation of 0.97 with custom capture data from the same samples (Pearson's r, P = 1.5 x 10^{-13}, Fig. 2.30B). Therefore, INVAR is not only highly sensitive when applied to custom capture panels that redundantly sequence up to 10^2-10^3 haploid genomes, but also when applied to WES data with a de-duplicated coverage between 10-100x (Fig. 2.30C).

We hypothesised that ctDNA could be detected and quantified with INVAR from even smaller amounts of input data. Therefore, we performed WGS on libraries from cfDNA of longitudinal plasma samples from a subset of six patients with Stage IV melanoma, to a mean depth of 0.6x (indicated in black in Fig. 2.28B). For each of those patients we identified >500 patient-specific mutations using WES from each patient's tumour and buffy coat DNA, generating between 226 and 7,696 IR per sample (median 861, IQR 471-1,559; Fig. 2.28B) with a "minimum family size" requirement of 1 (i.e. duplicate removal). Despite not leveraging unique molecular barcodes, error rates per trinucleotide were still sufficiently low, with many trinucleotide contexts showing error rates below 1 x 10^{-5} (Fig. 2.29). Using INVAR on sWGS data, IMAF values as low as 1.1 x 10^{-3} were quantified (Fig. 2.30D).

Compared to custom capture data from the same samples we observed a correlation of 0.93 (Pearson's r, P = 9 x 10^{-10}, Fig. 2.30B). In samples where ctDNA was not detected, it was possible to estimate the maximum likely IMAF of that sample from the known number of informative reads for each sample, which is indicated by the grey bars in Fig. 2.30D. Using less than 1x coverage, INVAR can boost the sensitivity utilising patient-specific mutation lists by up to an order of magnitude compared to copy-number analyses [27, 75].

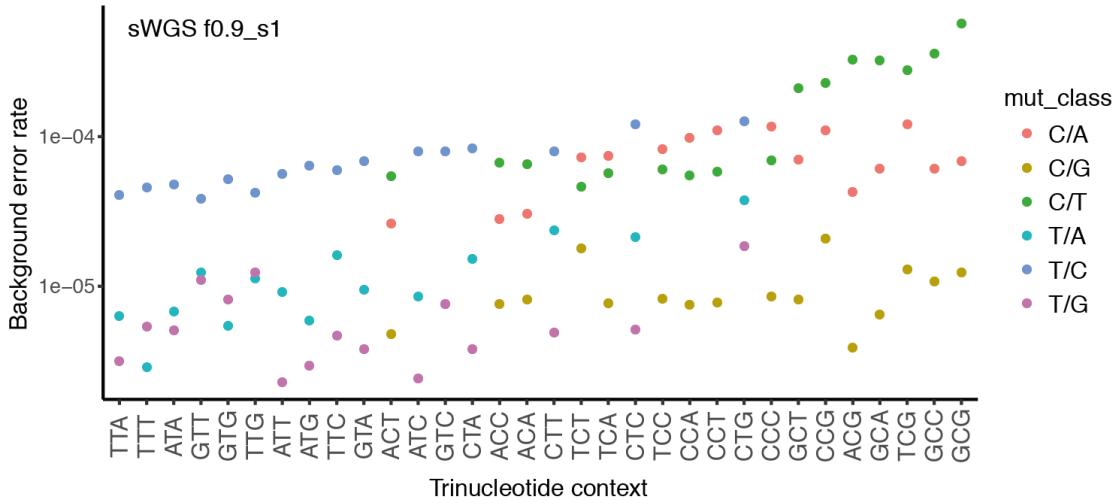


Fig. 2.29 **Application of INVAR to whole genome sequencing data.**
Background error rates per trinucleotide context for sWGS data. Data was analysed with INVAR using a minimum family size of 1 when performing read collapsing.

These analyses suggest that with a sufficiently large number of tumour -specific mutations, INVAR may yield high sensitivity for ctDNA detection from untargeted sequencing data that can be limited in depth and thus input material.

2.14 Extrapolation to higher IR and sensitivity

The sensitivity of INVAR depends on the number of patient-specific mutations identified, and so its effectiveness may be limited in samples with fewer identified mutations, or in cancers with lower mutation rates. Fig. 2.31A shows the distribution of IR for all the samples in this study, highlighting those with limited sensitivity (<20,000 IR) and those with sensitivity to ppm. Samples with limited sensitivity could be re-analysed with larger amounts of DNA input/more sequencing, or by designing larger capture panels by identifying additional patient specific mutations through broader-scale sequencing such as WGS of the tumour and

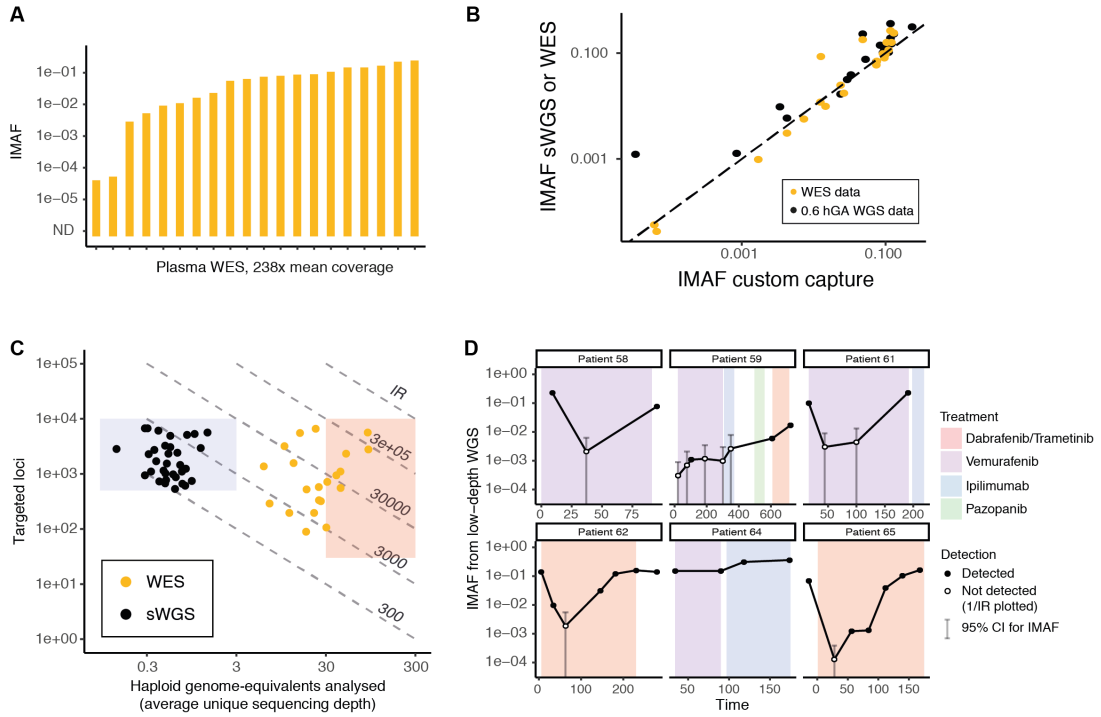


Fig. 2.30 **Sensitive detection of ctDNA from WES/WGS data using INVAR.** (**A**) IMAF observed for the 21 samples analysed with WES ordered from low to high. ND, not detected. (**B**) IMAFs obtained from plasma WES (gold) and sWGS (black) were compared to the IMAF obtained from the custom capture approach of matched samples, showing correlations of 0.97 and 0.93 (Pearson's r, P = 1.5 x 10^{-13} and P = 9 x 10^{-10}). (**C**) Number of hGA (indicating depth of unique coverage after read collapsing) and mutations targeted by plasma WES and sWGS. Compared to the custom capture approach, both the WES and sWGS samples had fewer hGA and occupy a space further to the left in the two-dimensional space. (**D**) sWGS longitudinal monitoring of ctDNA levels in plasma. All six patients had >500 mutations based on WES tumour profiling. Filled circles indicate detection at a specificity level of >0.99. For other samples, the 95% confidence intervals of the ctDNA level are shown, based on the number of informative reads for each sample (empty circles and bars). ND, not detected.

buffy coat DNA from that patient. Analyses involving greater IR would render the current background error rates limiting and would therefore require greater error-suppression, such as duplex molecular barcodes [115]. However, when increasing the sensitivity of ctDNA beyond the ppm range, sequencing output may become the limiting factor.

We studied the ctDNA levels in the samples analysed with custom capture in order to estimate the levels of sensitivity that would be required for different clinical applications. We used IMAF values from the clinical samples and plotted the detection rates while varying the numbers of IR and the levels of sensitivity (Fig. 2.31B). In Stage IV melanoma patients at pre-treatment baseline time points, ctDNA was detected in 100% of cases using 10^5 IR, whereas up to two orders of magnitude greater sensitivity may be needed to detect ctDNA at high rates following treatment initiation (Fig. 2.27B). For the population we studied of Stage II-III melanoma patients who underwent surgery, we suggest that even analysis of 10^7 IR might not be sufficient to detect all patients who ultimately relapse.

2.15 Discussion

In this study, we developed a method for sensitive patient-specific monitoring of ctDNA that leverages the properties of patient-specific sequencing data. This approach mitigates sampling error through aggregation of mutant signal, which is first weighted based on the features of each read and mutation locus, and uses features of cfDNA aside from specific sequence alterations, such as fragment sizes and tumour allele fractions of each mutation. By aggregating signal across 10^2-10^4 mutated loci it is possible to detect <0.01 copies of a cancer genome, even when this represents few parts per million of the cfDNA in plasma, 1-2 orders of magnitude lower than previous studies [65, 98].

We show that INVAR can be applied not only to patient-specific capture panel data to quantify ctDNA to parts per million (Fig. 2.21), but also to exome sequencing and sWGS data (Fig. 2.28 and Fig. 2.30). Although these latter methods generated fewer informative reads, INVAR detected ctDNA to 50 ppm using WES, and to 0.1% mutant allele fraction using sWGS, over an order of magnitude more sensitive compared to previous methods based on copy-number analysis of sWGS [75, 76]. This level of sensitivity can only be achieved by targeting a sufficiently large number of patient-specific mutations, as increasing input mass alone would not be feasible to this extent (Fig. 2.5A). Therefore, we assessed whether INVAR would still retain sufficient sensitivity when applied to sWGS data from other cancer types with lower mutation rates than melanoma. We estimated the potential sensitivity for INVAR using sWGS on other cancer types based on their known whole-genome mutation rates [116]

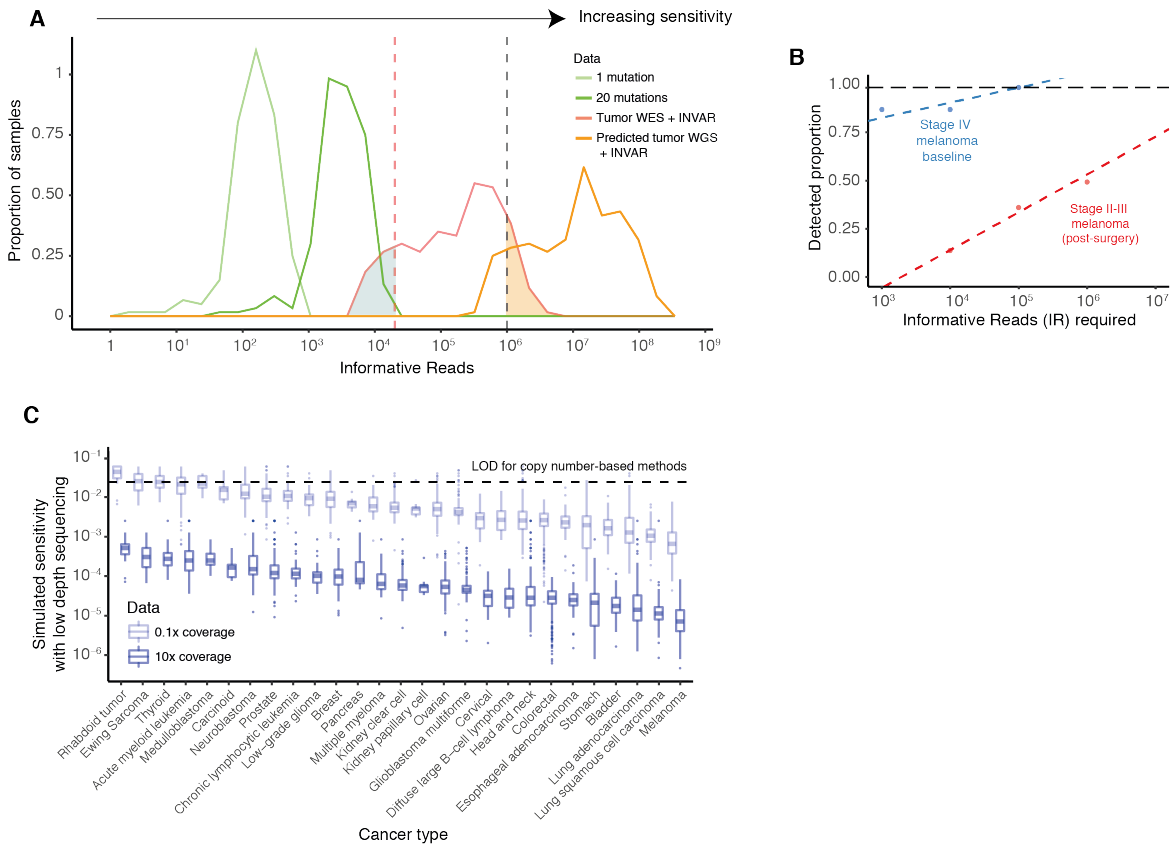


Fig. 2.31 **Current limitations and future applications of INVAR.**
(**A**) The number of informative reads that would be obtainable with different numbers of mutations analysed, across the cases in these three cohorts. Increasing sensitivity is directly correlated to IR, with the minimal detected ctDNA fraction being 2/IR in the current implementation of INVAR (section 2.16). The red line shows the distribution of IR obtained with the custom panels covering all mutations identified by tumour WES. Light/dark green lines show the IR generated if 1 or 20 mutations were analysed for each sample (calculated based on the mean IR per locus). IR could be increased further by using whole genome sequencing (WGS) to guide the design of custom panels (orange curve, extrapolated based on our observed mutation rates in WES). Using mutation lists from WES, samples exceeding 10^6 IR are shaded in orange, and samples with fewer than $2x10^4$ IR are shaded in blue. (**B**) Detection rates of ctDNA for different numbers of IR sequenced were estimated. There was a linear relationship between IR and detection ($R^2 = 0.95$) in the baseline samples of the stage IV melanoma cohort (blue). In stage II-III melanoma post-surgery (red), a linear relationship was observed between IR and detection rate, and the predicted rates of detection of ctDNA was extrapolated. ND, not detected. (**C**) Predicted sensitivities for sWGS plasma analysis of patients with different cancer types, using an average of 0.1x or 10x coverage (equivalent to 0.1 and 10 hGA) and the known mutation rates per Mbp of the genome for different patients and cancer types obtained from Lawrence et al. [116]. The limit of detection for ctDNA based on copy number alterations is shown at 3% [75].

(Fig. 2.31C). Using 0.1x WGS coverage, INVAR may yield sensitivities of 10^{-1}–10^{-3} for these cancer types, with the potential for higher sensitivity with deeper sequencing.

In summary, patient-specific mutation lists provide an opportunity for highly sensitive monitoring from a range of sequencing data types using methods for signal aggregation, weighting and error-suppression. As tumour sequencing becomes increasingly performed in personalised oncology, patient-specific mutation lists may be leveraged for individualised monitoring using INVAR-like tools.

2.16 Methods

2.16.1 Patient cohort

Samples were collected from patients enrolled on the MelResist (REC 11/NE/0312) and AVAST-M studies (REC 07/Q1606/15, ISRCTN81261306) [117, 118]. Consent to enter the studies was taken by a research/specialist nurse or clinician who was fully trained regarding the research. MelResist is a translational study of response and resistance mechanisms to systemic therapies of melanoma, including BRAF targeted therapy and immunotherapy, in patients with stage IV melanoma. AVAST-M is a randomised control trial which assessed the efficacy of bevacizumab in patients with stage IIB-III melanoma at risk of relapse following surgery; only patients from the observation arm were selected for this analysis. The Cambridge Cancer Trials Unit-Cancer Theme coordinated both studies, and demographics and clinical outcomes were collected prospectively. Baseline characteristics for all cohorts are summarised in Table 2.6.

2.16.2 Sample collection and processing

Fresh frozen tumour biopsies prior to treatment were collected from patients with Stage IV cutaneous melanoma. Formalin-fixed paraffin-embedded (FFPE) tumour tissue was obtained for the AVAST-M trial. For patients on the AVAST-M study, plasma samples were collected within 12 weeks of tumour resection, with a subsequent sample after 3 months, where available. Longitudinal samples were collected during treatment of patients with stage IV melanoma as part of the MelResist study. Peripheral blood samples were collected at each clinic visit in S-Monovette 9mL EDTA tubes. For plasma collection, samples were centrifuged at 1,600 g for 10 minutes within an hour of the blood draw, and then an additional centrifugation of 20,000 g for 10 minutes was carried out. All aliquots were stored at -80°C.

2.16.3 Tissue and plasma extraction and quantification

FFPE samples were sectioned into up to 8 µm sections, and one H&E stained slide was generated, which was outlined for tumour regions by a histopathologist. Marked tumour regions were macrodissected, and DNA extraction was performed using the QIAamp DNA FFPE Tissue Kit using the standard protocol, except with incubation at 56°C overnight and 500 rpm agitation on a heat block. DNA was eluted twice using 20 µL ATE buffer each time with centrifugation at full speed. Following extraction, DNA repair was performed using the NEBNext® FFPE DNA Repair Mix as per the manufacturer's protocol. Fresh frozen tissue biopsies were first homogenised prior to DNA extraction, which was performed as follows: up to 30 mg of each fresh frozen tissue biopsy sample was combined with 600 µL RLT buffer, then placed in a Precellys CD14 tube (Bertin Technologies) and homogenised at 6,500 rpm for two bursts of 20 seconds separated by 5 seconds. Subsequently, the Qiagen AllPrep extraction kit as per the manufacturer's protocol.

Genomic DNA was extracted from up to 1 mL whole blood or buffy coat using the Gentra Puregene Blood Kit (Qiagen) as per the manufacturer's protocol. Samples were eluted in two rounds of 70 µL buffer AE and incubated for 3 minutes before centrifugation. Up to 4mL of plasma was extracted using the QIAsymphony (Qiagen) with a QIAamp protocol. DNA was eluted in 90 µL elution buffer and stored at -80°C. Plasma samples were extracted using the QIAsymphony instrument (Qiagen) using the 2-4mL QIAamp protocol. For each QIAsymphony batch, 24 samples were extracted, which included a positive and negative control.

Following extraction of fresh frozen, FFPE and genomic DNA, eluted DNA concentration was quantified using a Qubit fluorimeter with a dsDNA broad range assay (ThermoFisher Scientific). To quantify cell-free DNA concentration of plasma DNA eluates, digital PCR was carried out using a Biomark HD (Fluidigm) with a 65bp TaqMan assay for the housekeeping gene RPP30 (Sigma Aldrich) [35]. 55 PCR cycles were used. The estimated number of RPP30 DNA copies per µL of eluate was used to determine the cell-free DNA concentration in the original sample.

2.16.4 Tumour library preparation

FFPE tumour tissue DNA samples (up to 150 ng) and buffy coat DNA samples (75 ng) were sheared to a length of 150bp, using the Covaris LE 220 (Covaris, Massachusetts, USA). The standard Covaris protocol for a final fragment length of 150bp and an input volume of 15µL using the 8 microTUBE-15 AFA Beads Strip V2 was used. After the shearing, the fragmentation pattern was verified using a Bioanalyser (Agilent).

Sequencing libraries were prepared using the ThruPLEX DNA-seq kit (Rubicon). 100 ng and 50 ng sheared tumour and buffy coat DNA, respectively, were used and the protocol was carried out according to the manufacturer's instructions. The number of amplification cycles was varied during library preparation according to the manufacturer's recommendations. Library concentration was determined using qPCR with the Illumina/ROX low Library Quantification kit (Roche). Library fragment sizes were determined using a Bioanalyser (Agilent). After library preparation, exome capture was performed with the TruSeq Exome Library Kit (Illumina), using a 45Mbp exome baitset. Three libraries were multiplexed in one capture reaction and 250 ng of each library was used as input. For compatibility with ThruPLEX libraries, the protocol was altered by adding 1μL of i5 and i7 TruSeq HT xGen universal blocking oligos (IDT) during each hybridisation step. To compensate for the increased hybridisation volume, the volume of CT3 buffer was adjusted to 51 μL. Two rounds of hybridisations were carried out, each lasting for 24 hours. Library QC was performed using qPCR and Bioanalyser, as above. Samples were multiplexed and sequenced with a HiSeq 4000 (Illumina).

Fresh frozen tumour biopsies and matched buffy coat library preparation was performed as described by Varela et al. [103] using the SureSelectXT Human All Exon 50 Mb (Agilent) bait set. Samples were multiplexed and sequenced with a HiSeq 2000 (Illumina).

2.16.5 Tumour mutation calling

For fresh frozen tumour biopsies, mutation calling was performed as described by Varela et al. [103]. For FFPE tumour biopsies, mutation calling was performed with Mutect2 with the default settings: –cosmic v77/cosmic.vcf and –dbsnp v147/dbsnp.vcf. To maximize the number of mutations retained, variants achieving Mutect2 pass OR tumour LOD > 5.3 were retained. Mutation calls were filtered as follows:

1. Buffy coat mutant allele fraction equals zero
2. Mutation not in homologous region
3. Mutation not at a multiallelic locus
4. 1000 Genomes ALL and EUR frequency equals zero
5. A minimum unique tumour depth of 5.

In addition, for FFPE based tissue data in the melanoma cohort, the filter for C/A errors proposed by Costello et al. [112] was applied to suppress C/A artefacts. The remaining

mutations are shown in Fig. 2.9. The remaining signal in the C/A class, especially within the trinucleotide contexts CCC and CCA indicate that artefactual mutations still remain in the data. Additonal filtering in these contexts, for example by allelic fraction, should be conducted to remove further artefactual mutations [112]. As a result, we generated patient-specific mutation lists for 47 patients with stage II-IV melanoma. A median of 625 (IQR 411 -1076) patient-specific mutations were identified per patient (Fig. 2.8, Table 2.1). These mutation lists were used both to design custom capture sequencing panels, and as input for the INVAR method.

2.16.6 Plasma library preparation

Cell-free DNA samples were vacuum concentrated at 30°C using a SpeedVac (ThemoFisher) prior to library preparation where required. The median input into the library was 1652 haploid genomes (IQR 900 – 3013). Whole genome library preparation for plasma cell-free DNA was performed using the Rubicon ThruPLEX Tag-Seq kit. The number of PCR amplification cycles during the ThruPLEX protocol was varied between 7-15 cycles, as recommended by the manufacturer. Following amplification and sample barcoding, libraries were purified using AMPure XP beads (Beckman Coulter) at a 1:1 ratio. Library concentration was determined using the Illumina/ROX low Library Quantification kit (Roche). Library fragment sizes were determined using a Bioanalyser (Agilent).

For the stage IV melanoma cohort, library preparation and sequencing were run in duplicate to assess the technical reproducibility of the experimental and computational method, showing a correlation between IMAF values generated by the INVAR pipeline of 0.97 (Pearson's r, p-value < 2.2 x 10^{-16}). For the early-stage cohorts, input cell-free DNA material was not split and was instead prepared and sequenced as a single sample per time point.

2.16.7 Custom hybrid-capture panel design and plasma sequencing

Following mutation calling, custom hybrid-capture sequencing panels were designed using the Agilent SureDesign software. Between 5 and 9 patients were grouped together per panel in this implementation. Baits were designed with 4-5x density and balanced boosting. 95.5% of the variants had baits successfully designed; bait design was not reattempted for loci that had failed. Custom panels ranged in size between 1.26-2.14 Mb with 120 bp RNA baits. For each panel, mutation classes and tumour allele fractions are shown in Fig. 2.9, Fig. 2.10 and Table 2.1.

Libraries were captured either in single or 3-plex (to a total of 1000 ng capture input) using the Agilent SureSelectXT protocol, with the addition of i5 and i7 blocking oligos (IDT) as recommended by the manufacturer for compatibility with ThruPLEX libraries [119]. Custom Agilent SureSelectXT baits were used, with 13 cycles of post-capture amplification. Post-capture libraries were purified with AMPure XP beads at a 1:1.8 ratio, then were quantified and library fragment size was determined using a Bioanalyser (Agilent).

2.16.8 Exome capture sequencing of plasma

For exome sequencing of plasma, the Illumina TruSeq Exome capture protocol was followed. Libraries generated using the Rubicon ThruPLEX protocol (as above) were pooled in 3-plex, with 250 ng input for each library. Libraries underwent two rounds of hybridisation and capture in accordance with the protocol, with the addition of i5 and i7 blocking oligos (IDT) as recommended by the manufacturer for compatibility with ThruPLEX libraries. Following target enrichment, products were amplified with 8 rounds of PCR and purified using AMPure XP beads prior to QC.

2.16.9 Plasma sequencing data processing

Cutadapt v1.9.1 was used to remove known 5' and 3' adaptor sequences specified in a separate FASTA of adaptor sequences. Trimmed FASTQ files were aligned to the UCSC hg19 genome using BWA-mem v0.7.13 with a seed length of 19. Error-suppression was carried out on ThruPLEX Tag-seq library BAM files using CONNOR [67]. The consensus frequency threshold -f was set as 0.9 (90%), and the minimum family size threshold -s was varied between 2 and 5 for characterisation of error rates. For custom capture and exome sequencing data, a minimum family size of 2 was used. For sWGS a minimum family size of 1 was used, i.e. not using molecular barcodes except for where duplicates are present.

To leverage signal across multiple time points, error-suppressed BAM files could be combined using 'samtools view -ubS - | samtools sort -' prior to further data processing. In the early-stage melanoma cohort (AVAST-M), where multiple samples were available for the same patient before 6 months post-surgery, sequencing data for each of the samples were merged.

2.16.10 Low-depth whole-genome sequencing of plasma

For sWGS, 30 libraries were sequenced per lane of HiSeq 4000, achieving a median of 0.6x deduplicated coverage per sample. For these libraries, since the number of informative

reads (IR) would limit sensitivity before background errors would become limiting, we used error-suppression with family size 1 for this particular setting. Error rates per trinucleotide were compared between WGS and custom hybrid-capture sequencing data for family size 1, showing a Pearson r of 0.91. WGS data underwent data processing (section 2.17) except the minimum depth at a locus was set to 1, and patient-specific outlier-suppression (section 2.17) was not used because loci with signal vs. loci without signal would only give allele fractions of 0 or 1 given a depth of 0.6x.

2.16.11 INVAR pipeline

The INVAR pipeline takes BAM files (+/- error-suppression with molecular barcodes), a BED file of patient-specific loci, and a CSV file indicating the tumour allele fraction of each mutation and which patient it belongs to. The pipeline is shown in Fig. 2.7 and full details are given in the section 2.17. See 'Data and materials availability' for code access.

2.16.12 Imaging

CT imaging was acquired as part of the standard of care from each patient of the stage IV melanoma cohort and was examined retrospectively. Slice thickness was 5 mm in all cases. All lesions with a diameter greater than 5 mm were outlined slice by slice on CT images by an experienced operator, under the guidance of a radiologist, using custom software written in MATLAB (Mathworks, Natick, MA). The outlines were subsequently imported into the LIFEx software [120] in NifTI format for processing. Tumour volume was then reported by LIFEx as an output parameter from its texture-based processing module (Table 2.5).

2.16.13 Data and materials availability

Raw sequencing data will be made available at the European Genome-phenome archive, accession number EGAS00001002959. The INVAR pipeline will be made publicly accessible at http://www.bitbucket.org/nrlab/invar.

2.17 Supplementary methods

2.17.1 INVAR data processing

SAMtools mpileup 1.3.1 was used at patient-specific loci based on a BED file of mutations, with the following settings: –ff UNMAP, –q 40 (mapping quality), –Q 20 (base quality), –x,

–d 10,000, then multiallelic calls were split using BCFtools 1.3.1. Next, all TSV files were annotated with 1,000 Genomes SNP data, COSMIC data, and trinucleotide context using a custom Python script. Output files were then concatenated, compressed, and read into R. First, based on prior knowledge from tumour sequencing data, all loci were annotated per patient with being either: patient-specific (present in patient's tumour) or non-patient-specific (not present in patient's tumour, or individual does not have cancer). Data points were excluded if MQSB < 0.01 (mapping quality / strand bias). Since each non-patient-specific sample contains the loci from multiple patients, every non-patient-specific sample may control for all other patients analysed with the same sequencing panel or method (excluding loci that are shared between individuals).

2.17.2 INVAR data filters I

The following filters were applied to INVAR data:

1. Loci that showed mutant signal in >10% of the non-patient-specific (patient-control) samples were blacklisted. For custom capture and exome sequencing data, we required the mean background error rate in each locus to be <1% mutant allele fraction, otherwise the locus was classed as noisy. The proportion of loci that were blacklisted with these filters ranged from 0.21%-3.53% (Fig. 2.13D). Patient samples may be used to characterise the noise per locus (at loci that did not belong to them), since 99.8% of mutations were private to each patient. Control sample QC data are shown in Table 2.3.

2. Mutation signal had to be represented in both the F and R read of that read pair (Fig. 2.13C). The resulting error-suppression is analogous to tools that merge paired-end reads [121].

2.17.3 INVAR data annotation

After data filtering, data was annotated with both locus noise filter and trinucleotide error rate. Since the locus noise filter is limited by the number of control samples and cfDNA molecules at that locus, we also assessed trinucleotide error rate. Trinucleotide error rates were determined from the region up to 10bp either side of every patient-specific locus (excluding patient-specific locus itself), and data was pooled by trinucleotide context. After pooling data in this manner, a median of 3.0 x 10^8 informative reads (or deduplicated reads) per trinucleotide context were analysed. Trinucleotide error rate was calculated as a mismatch rate for each specific mutation context. If a trinucleotide context had zero mutant

deduplicated reads, the error rate was set to the reciprocal of the number of IR/deduplicated reads in that context.

In addition, each data point was annotated with the cfDNA fragment size of that read using a custom Python script. Then, to eliminate outlier signal that was not consistent with the remainder of that patient's loci, we performed patient-specific outlier suppression (Fig. 2.15B and C, Fig. 2.14C). The data is now error-suppressed (both by read-collapsing and bespoke methods for patient-specific sequencing data) and annotated with parameters required for signal-enrichment (by features of ctDNA sequencing) for the INVAR method.

2.17.4 INVAR data filters II - patient-specific outlier-suppression

Patient-specific sequencing data consists of informative reads at multiple known patient-specific loci, providing the opportunity to compare mutant allele fractions across loci as a means of error-suppression. The distribution of signal across loci potentially allows for the identification of noisy loci not consistent with the overall signal distribution. Each locus was tested for the probability of having observed mutant reads given the average signal across all loci (Fig. 2.15B and C, Fig. 2.14C). A locus observed with significantly greater signal than the remainder of the loci might be attributed to noise at that locus, contamination, or a mis-genotyped SNP locus. The possibility of a mis-genotyped SNP becomes increasingly likely when a large number of mutated loci are targeted by INVAR.

For each sample, the IMAF was determined across all loci passing pre-INVAR data processing filters with mutant allele fraction at that locus of <0.25, similar to that outlined by Phallen et al. [64], who used a similar threshold, and Abbosh et al. [19], who used a threshold of 0.20. Loci with signal >0.25 mutant allele fraction were not included in the calculation because (i) in the residual disease setting, loci would not be expected to have such high mutant allele fractions (unless they were mis-genotyped SNPs), and (ii) if the true IMAF of a sample is >0.25, when a large number of loci are tested, they will show a distribution of allele fractions such that detection is supported by having many low allele fraction loci with signal.

Based on the ctDNA level of the sample, the binomial probability of observing each individual locus given the IMAF of that sample was calculated. Loci with a Bonferroni corrected P-value <0.05 (corrected for the number of loci interrogated) were excluded in that sample, thereby suppressing outliers. As a result of outlier-suppression, background noise was reduced to 33% in control samples, while retaining 96.1% of signal in patient samples (Fig. 2.15C). By correcting the P-value threshold for the number of loci tested, this filter can be applied to data with a variable number of mutations targeted per patient, enabling analysis of samples from patients with cancer types with both high and low mutation rates.

2.17.5 Statistical detection method for INVAR

We developed a statistical method to model the number of mutant reads at multiple patient-specific loci, incorporating prior information available from patient-specific sequencing, such as the background error of the trinucleotide context, the tumour allele fraction at the locus, and fragment length. This approach aggregates signal across multiple patient-specific mutations following error-suppression. For each locus, we test the significance of the number of mutant reads given the trinucleotide error rate of that context. Trinucleotide error rates were used instead of locus-specific error rates in order to determine a more accurate estimation of background error rates to 10^{-7} (Fig.2.14A).

Tumour allele fractions and trinucleotide error rates were considered as follows: Denote AF_i as the tumour mutant allele fraction at locus i, e_i as the background error in the context of locus i, and let p be an estimate of ctDNA content in that sample for the INVAR pipeline. A random read at locus i can be observed to be mutant either if it arose from a mutant molecule, or an incorrectly sequenced wild type DNA molecule. This occurs with probability q_i:

$$q_i = AF_i \cdot (1 - e_i) \cdot p + (1 - AF_i) \cdot e_i \cdot p + e_i \cdot (1 - p) \tag{1}$$

Testing for the presence of ctDNA is now equivalent to testing the statistical hypothesis H_0: $p = 0$. Assuming the number of observed mutant reads is independent between loci, the following likelihood function can be produced:

$$L(p; M, AF, e) = \prod_{i=1}^{n} \prod_{j=1}^{R_i} q_i^{M_{ij}} (1 - q_i)^{1 - M_{ij}} \tag{2}$$

where M_{ij} is the indicator for a mutation in read j of locus i, and R_i is the number of reads in locus i. The above method allows weighting of signal by tumour allele fraction, which we confirm influences plasma mutation representation in patient samples with early stage and advanced disease (Fig. 2.17A), and in the spike-in dilution series from one patient (Fig. 2.16A).

Each sequencing read provides fragment size information (Fig. 2.16B), which may be used to separate mutant from wild-type molecules and produce an enrichment in ctDNA

(Fig. 2.17B). Probability weighing was preferred over size selection to avoid allelic loss at ultra-low allele fractions, suggested by Fan et al. [12] in the non-invasive prenatal testing setting. Therefore, read length information can also be incorporated into the likelihood. The method for read length distribution of mutant and wild-type fragments estimation is given in section 2.17.9. This approach is in contrast to size-selection and may be considered as a size-weighting step alongside tumour AF weighting that was performed above. Fragment sizes for each sequencing read may be incorporated to the INVAR method. To do so, let L_{ji} be the length of read j at locus i. The likelihood can be written as:

$$L(p;M,L,AF,e) = \prod_{i=1}^{n}\prod_{j=1}^{R_i} P(m_{ij},l_{ij}|e,AF,p) \tag{3}$$

Assuming that given the read length and mutation status are independent given the source of the read (mutant or wild-type DNA), we can factor the likelihood as follows:

$$L(p;M,L,AF,e) =$$

$$\prod_{i=1}^{n}\prod_{j=1}^{R_i} P(m_{ij},l_{ij}|z_{ij}=0)\cdot P(z_{ij}=0)+P(m_{ij},l_{ij}|z_{ij}=1)\cdot P(z_{ij}=1) =$$

$$\prod_{i=1}^{n}\prod_{j=1}^{R_i} P(m_{ij}|z_{ij}=0)\cdot p^0(l_{ij})\cdot(1-p)+P(m_{ij}|z_{ij}=1)\cdot p^1(l_{ij})\cdot p =$$

$$\prod_{i=1}^{n}\prod_{j=1}^{R_i} e_i^{m_{ij}}\cdot(1-e_i)^{1-m_{ij}}\cdot p^0(l_{ij})\cdot(1-p)+g_i^{m_{ij}}\cdot(1-g_i)^{1-m_{ij}}\cdot p^1(l_{ij})\cdot p \tag{4}$$

where z_{ij} is the indicator that read j of locus i came from ctDNA, $p^k(l_{ij}) = P(l_{ij} \mid z_{ij} = k)$, and $g_i = AF_i\cdot(1-e_i)+(1-AF_i)\cdot e_i$. The above method weights the signal based on both fragment length of mutant and wild-type reads, though in this implementation of INVAR, we set the weight of all wild-type size bins to be equal, thereby neglecting size information from wild-type reads.

Lastly, a score is generated for each sample through aggregation of signal across all patient-specific loci in that sample using the Generalised Likelihood Ratio test (GLRT). The GLRT directly compares the likelihood under the null hypothesis against the likelihood under the maximum likelihood estimate of p:

$$\lambda(p_0) = \frac{L(p_0; M, L, AF, e)}{L(\hat{p}; M, L, AF, e)} \tag{5}$$

The higher the value of the likelihood ratio, the greater evidence for ctDNA presence in a sample. Classification of samples was performed based on comparison of likelihood ratios between patient and control samples.

2.17.6 Likelihood ratio threshold determination

Other patients were used to control for one another at non-shared loci (Fig. 2.12D). Only samples run on the same sequencing panel (i.e. same custom sequencing panel design), with the same error-suppression setting and targeting the same mutation list were used to control for one another.

In order to determine an accurate threshold for the likelihood ratio (LR) based on controls, reads from each control sample were iteratively resampled with replacement 10 times, and the GLRT script was run. To minimise the risk of any patient-specific contamination of signal at non-patient-specific control loci (through de novo mutations overlapping with patient-specific sites), only samples with patient-specific IMAF <1% were used as controls for determination of the cut-point. Control samples were required to have at least as many IR as the patient sample with the fewest IR in that cohort, otherwise they were excluded.

Based on the LR distribution in patient controls and patient samples, the cut-off for LR was determined for each cohort using the 'OptimalCutpoints' package in R [122], maximising sensitivity and specificity using the 'MaxSnSp' setting. Based on the LRs per cohort, an analytical specificity was determined for each cohort (Fig. 2.20, Table 2.4).

2.17.7 Assessment of specificity in healthy individuals

26 healthy individuals' cfDNA from plasma were analysed using the stage IV melanoma cohort. These samples were treated as 'patient' samples, and so had no influence on the

filters in the pipeline and were not used for the determination of LR thresholds. After determination of the LR thresholds (described above), the LRs from healthy individuals' samples were assessed for false positive detection of ctDNA. For each of the INVAR applications (custom capture, WES and sWGS), the clinical specificity values in healthy individuals were determined (Fig. 2.20, Table 2.4).

2.17.8 Estimation of ctDNA content per sample for likelihood ratio determination

In this section we derive an Expectation Maximisation (EM) algorithm to estimate p as part of the INVAR method. If we treat the tumour of origin z_{ij} as a latent variable, and assume that it is known, the joint likelihood of Z, M (m_{ij} is the indicator for a mutation in read j of locus i), L (l_{ij} is the length of read j of locus i), AF (AF_i is the tumour allele fraction at locus i), e (e_i is the background error in the context of locus i) can be written as:

$$L(p;Z,M,L,AF,e)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{R_i} [e_i m^{ij}] \cdot p^0(l_{ij}) \cdot (1-p)]^{1-z_{ij}}$$

$$\cdot [g_i^{mij} \cdot (1-AF_i) \cdot p^1(l_{ij}) \cdot p]^{z_{ij}}$$

Where $g_i = AF_i \cdot (1-e_i) + (1-AF_i) \cdot e_i$. The log-likelihood is linear in z_{ij}, so taking the expectation of the likelihood amounts simply to replacing the z_{ij} with their expectation at stage l, $z_{ij}^l = E(z_{ij} \mid m_{ij}, l_{ij}, p_l)$, where p_l is the best estimate of p at iteration l. We can thus use EM to find a maximum likelihood estimate for p, by iteratively maximising the likelihood with respect to p, and taking the expectation of the likelihood with respect to z_{ij}. An estimate for p_l is obtained by taking the derivative with respect to p_l and equating it to zero:

$$p_l = \frac{\sum_{i=1}^{n} \sum_{j=1}^{R_i} z_{ij}^l}{\sum_{i=1}^{n} R_i}$$

The above is simply the expected proportion of reads from ctDNA at stage l. Bayes' theorem can be used to compute z_{ij}^l:

$$z_{ij}^l = P(z_{ij} = 1 | m_{ij}, l_{ij}, p_l)$$

$$= \frac{P(m_{ij}|z_{ij}=1)\cdot p^1(l_{ij})\cdot p}{P(m_{ij}|z_{ij}=1)\cdot p^1(l_{ij})\cdot p + P(m_{ij}|z_{ij}=0)\cdot p^0(l_{ij})\cdot(1-p))}.$$

By substituting the respective probabilities we obtain:

$$z^l_{ij} = \frac{g_i^{m_{ij}}(1-g_i)^{1-m_{ij}}p^1(l_{ij})p}{g_i^{m_i j}(1-g_i)^{1-m_{ij}}p^1(l_{ij})p + e_i^{m_{ij}}(1-e_i)^{1-m_{ij}}p^0(l_{ij})(1-p)}$$

The algorithm proceeds by alternating the maximisation of p, and the expectation of the z_{ij}.

2.17.9 Estimation of read length distribution for INVAR

Size-weighting with INVAR depends on first having a known distribution of sizes of mutant and wild-type reads against which to perform weighting. In order to estimate the read length distribution with the greatest accuracy, we used all wild type and mutant reads from all samples in that panel, leaving out the sample being tested (i.e. leave-one-out approach), and we used kernel density estimation to smooth the respective probabilities.

The size distributions from each of the studied cohorts are shown in Fig. 2.16, and the enrichment ratios for each size range are shown in Fig. 2.17B. We demonstrated that the early stage melanoma cohort had a significantly different size profile from the advanced stage melanoma cohort, which had a significantly greater proportion of di-nucleosomal fragments despite downsampling of data to a similar number of reads (Fig. 2.16C). Thus, the fragment length distribution of mutant and wildtype fragments was assessed separately in the two cohorts, and data was smoothed with a Gaussian kernel with a default setting of 0.25 (Fig. 2.16D).

To estimate the probability that a read is of length l, given that the cell of origin is wild type, $P(L = l \mid z = 0)$, we used all of the wild-type reads from each pooled dataset. For both of the data sets, we used the R function "density", with a Gaussian kernel, to smooth the estimated probabilities, and obtained a density estimate $\hat{f}(l \mid Z = z)$. Finally, to estimate $P(L = l \mid z = z)$, we integrated the respective density:

$$P(L = l \mid Z = z) = \int_{l-0.5}^{l+0.5} \hat{f}(t \mid Z = z)dt$$

Smoothing the size distribution estimates is important in datasets where data is sparse to avoid assigning too large a weight to any given mutant fragment.

2.17.10 Calculation of informative reads (IR)

The number of informative reads (IR) for a sample is the product of the number of mutations targeted (i.e. length of the mutation list) and the number of haploid genomes analysed by sequencing (hGA, equivalent to the deduplicated coverage following read-collapsing). Thus, the limit of detection for every sample can be calculated based on 1/IR (with adjustment for sampling mutant molecules based on binomial probabilities). For non-detected samples, the 1/IR value provides an estimate for the upper limit of ctDNA in that sample; this allows quantification of samples even if no mutant molecules are present, and is utilised in Fig. 2.30D to define the upper confidence limits to $\sim 10^{-4}$ using sWGS data. Also, samples with limited sensitivity can be identified and classified as a 'low-sensitivity' or 'non-evaluable' group, where the INVAR method is limited by the number of IR (Fig. 2.31A). In this study, we aimed to quantify ctDNA with sensitivity greater than other methods, and classified samples with non-detected ctDNA with <20,000 IR as low-sensitivity and thus non-evaluable. Across the cohorts in this study, 4 patients were non-evaluable with these criteria.

2.17.11 Calculation of integrated mutant allele fraction (IMAF)

To quantify ctDNA across multiple mutated loci, we calculated an 'integrated mutant allele fraction', as follows:

- For each trinucleotide context in a sample, the deduplicated depth-weighted mean allele fraction across all patient-specific loci was calculated
- The background error rate per trinucleotide context in control data was subtracted from the mean allele fraction calculated in (a). Trinucleotide contexts with negative mutant allele fraction after subtraction were set to zero.
- The mean background-subtracted allele fraction was taken across the trinucleotide contexts, weighted by the deduplicated depth in each trinucleotide context.

2.17.12 Experimental spike-in dilution series

Plasma DNA from one patient with a total of 5,073 patient-specific variants was serially diluted 10-fold each step in a pool of plasma cfDNA from 11 healthy individuals (Seralab) to give a dilution series spanning 1-100,000x. Library preparation was performed, as described above, with 50 ng input per dilution. In order to interrogate a sufficiently large number of molecules in the dilution series to assess sensitivity, the lowest dilution (100,000x) was

generated in triplicate. The healthy control cfDNA pools were included as control samples for the determination of locus error rate to identify and exclude potential SNP loci (Fig. 2.19A).

Given the relationship between tumour allele fraction and plasma mutation representation (Fig. 2.16A), any smaller panel for INVAR should be based on clonal mutations with highest priority, with lower allele fractions included only if plasma sequencing data is sufficiently broad. Thus, we iteratively sampled the data with replacement from each of the dilution series sequencing libraries (with 50 iterations), and then selected the top N mutations (spanning 1 to 5,000 mutations). The locus with the highest mutant allele fraction was the BRAF V600E mutation. After downsampling the number of loci, outlier-suppression was repeated on all samples except for the single BRAF V600E locus data (Fig. 2.19B).

2.18 Supplementary tables

chr	pos	ref	alt	gene	depth	tumour_af	patient	study	cancer_type	panel_no
chr2	160801440	C	T	PLA2R1	8	0.99	12	AVASTM	melanoma	7
chr2	102851426	C	T	IL1RL2	6	0.99	1	AVASTM	melanoma	4
chr5	112176316	T	C	APC	7	0.99	1	AVASTM	melanoma	4
chr6	32610081	G	A	HLA-DQA1	9	0.99	1	AVASTM	melanoma	4
chr10	25861748	G	A	GPR158	5	0.99	1	AVASTM	melanoma	4
chr17	21117578	G	A	TMEM11	5	0.99	1	AVASTM	melanoma	4
chrX	3235566	C	T	MXRA5	9	0.99	10	AVASTM	melanoma	4
chr1	151687158	G	A	CELF3	8	0.99	30	AVASTM	melanoma	2
chr7	141791779	G	A	MGAM	5	0.99	33	AVASTM	melanoma	5
chr14	19563436	C	T	POTEG	5	0.99	33	AVASTM	melanoma	5
chr6	106555323	A	T	PRDM1	5	0.99	38	AVASTM	melanoma	6
chrX	30237036	G	A	MAGEB2	10	0.99	38	AVASTM	melanoma	6
chrX	105279276	C	T	SERPINA7	7	0.99	38	AVASTM	melanoma	6

Table 2.1 **Patient-specific mutation lists (selected indicative rows out of 55042 rows)**
This table contains all patient-specific mutation lists for patients in this study. The following cohorts are represented: AVASTM (stage II-III melanoma) and MELR (stage IV melanoma). Mutation positions are given using the hg19 genome build. chr, chromosome; pos, position; ref, reference allele; alt, alternate allele; depth, deduplicated tumour sequencing depth; tumour_af, allelic fraction in corresponding tumour tissue; patient, number of patient in study; study, study the patient belongs to; cancer_type, cancer type of the patient; panel_no, patient-specific mutation list number (multiple patients were grouped for panel design and analysis).

patient	TP	DP pre.dedup	lib input	data type	IMAF	INVAR LR	IR	mut sum	Detected	cancer genomes	targeted muts	hGA	LS PASS
58	4	0.86	10.04	sWGS	7.60E-02	1.49E+02	605	46	TRUE	4.88E-02	943	0.64	NA
58	1	0.73	10.06	sWGS	2.31E-01	4.17E+02	462	108	TRUE	1.15E-01	943	0.49	NA
59	22	0.82	10	sWGS	1.70E-02	6.27E+01	1528	26	TRUE	9.51E-03	2733	0.56	NA
59	1	1.67	9.4	sWGS	0	0	3337	0	FALSE	0	2733	1.22	NA
61	1	1.47	4.6	sWGS	1.02E-01	2.56E+02	760	79	TRUE	1.01E-01	784	0.97	NA
61	8	0.99	10.07	sWGS	2.32E-01	4.23E+02	471	112	TRUE	1.43E-01	784	0.6	NA
62	1	0.93	9.96	sWGS	1.41E-01	3.75E+02	700	100	TRUE	7.34E-02	1363	0.51	NA
62	9	1.48	10.06	sWGS	1.41E-01	5.49E+02	979	143	TRUE	1.05E-01	1363	0.72	NA
64	1	1.05	8.6	sWGS	1.52E-01	1.98E+02	369	57	TRUE	9.71E-02	587	0.63	NA
64	6	1.31	10	sWGS	3.63E-01	7.17E+02	483	180	TRUE	3.07E-01	587	0.82	NA
65	1	1.50	9.99	sWGS	6.87E-02	1.38E+03	5529	383	TRUE	7.20E-02	5317	1.04	NA
58	2	0.70	5.84	sWGS	0	0	476	0	FALSE	0	943	0.5	NA
59	18	0.75	8.91	sWGS	5.92E-03	1.54E+01	1345	8	TRUE	2.93E-03	2733	0.49	NA
61	2	0.74	8.72	sWGS	0	0	333	0	FALSE	0	784	0.42	NA

Table 2.2 **Sample library preparation input, QC, and INVAR likelihood ratios – test samples (selected indicative rows out of 201 rows)**
For all patient samples, QC metrics, ctDNA IMAF values and sequencing depths are listed. Samples were non-evaluable if they had no ctDNA signal and <20,000 IR. patient, patient number; TP, timepoint the sample was taken; DP pre.dedup, mean depth per patient before deduplication; lib input, ng input into library prep; data type, custom capture, WES or sWGS; IMAF, integrated mutant allele fraction; INVAR LR, likelihood ratio; IR, informative reads; mut sum, total sum of mutant collapsed reads observed in the sample; Detected, sample was detected positive for ctDNA using size weighting approach; cancer genomes, indicates the number of cancer genomes present in the sample, based on the number of mutant reads and the number of loci targeted; targeted muts, number of mutated loci targeted in the sample; hGA, haploid genomes analysed; LS PASS, indicating samples passing low sensitivity threshold (set at $>$ 20,000IR). This filter is not applicable to the whole genome sequencing cohort.

individual	DP_pre.dedup	lib_input	data_type	source_pt	IMAF	INVAR_LR	IR	mut_sum	Detected	targeted_mut	cancer_genomes	hGA
PPC7-1_85	374.13	28.4	WES	63	0	0	8145	0	FALSE	274	0	29.73
PPC7-1_85	374.13	28.4	WES	66	0	0	7928	0	FALSE	237	0	33.45
PPC7-1_85	374.13	28.4	WES	56	0	0	29268	0	FALSE	452	0	64.75
PPC7-1_85	374.13	28.4	WES	59	0	0	105748	0	FALSE	2733	0	38.69
PPC7-1_85	374.13	28.4	WES	64	0	0	20414	0	FALSE	587	0	34.78
PPC7-1_85	374.13	28.4	WES	62	0	0	39301	0	FALSE	1363	0	28.83
PPC7-1_85	374.13	28.4	WES	58	0	0	40913	0	FALSE	943	0	43.39
PPC7-1_85	374.13	28.4	WES	61	0	0	32485	0	FALSE	784	0	41.43
PPC7-1_85	374.13	28.4	WES	65	7.03E-06	0.55	229223	2	TRUE	5317	0.00038	43.11
PPC7-2_90	358.32	34.4	WES	63	0	0	7045	0	FALSE	274	0	25.71
PPC7-2_90	358.32	34.4	WES	56	0	0	25055	0	FALSE	452	0	55.43
PPC7-2_90	358.32	34.4	WES	59	0	0	94563	0	FALSE	2733	0	34.6
PPC7-2_90	358.32	34.4	WES	62	0	0	34774	0	FALSE	1363	0	25.51
PPC7-2_90	358.32	34.4	WES	65	0	0	199988	0	FALSE	5317	0	37.61

Table 2.3 **Sample library preparation input, QC, and INVAR likelihood ratios – control samples (selected indicative rows out of 202 rows)**

For all control samples, QC metrics, ctDNA IMAF values and sequencing depths are listed. Individual, sample name; DP_pre.dedup, mean depth per patient before deduplication; lib_input, ng input into library prep; data_type, custom capture, WES or sWGS; source_pt, patient the targeted mutation belongs to; IMAF, integrated mutant allele fraction; INVAR_LR, likelihood ratio score from the INVAR pipeline; IR, informative reads; mut_sum, total sum of mutant collapsed reads observed in the sample; Detected, sample was detected positive for ctDNA using size weighting approach; targeted_mut, number of mutated loci targeted in the sample; cancer_genomes, indicates the number of cancer genomes present in the sample, based on the number of mutant reads and the number of loci targeted; hGA, haploid genomes analysed.

study	LR_threshold	analytical_specificity	specificity	fam_size
WES_samples	3.19	96.18%	95.79%	2
MELR	6.62E-07	95.41%	96.74%	2
AVASTM_MERGED	4.43E-01	98.37%	NA	2
sWGS_samples	5.96	100%	97.60%	1

Table 2.4 **INVAR score thresholds**
This table gives details on each of the cohorts, the experimental method performed to generate data, and the INVAR score threshold used (determined by ROC analysis). LR_threshold, likelihood ratio threshold used for detection; analytical_specificity, determined using other patients as control samples; specificity, determined using healthy individuals on the same panel; fam_size, minimum family size setting used in read-collapsing.

patient	days	volume_mL	n_lesions	TP	evaluable
58	0	23.89	1	1	TRUE
58	63	11.37	1	3	TRUE
59	0	4.61	5	1	TRUE
59	75	1.02	5	3	TRUE
59	187	1.30	5	7	TRUE
59	300	5.49	5	11	TRUE
59	406	28.28	7	16	TRUE
59	609	157.49	7	18	TRUE
59	720	307.58	5	22	TRUE
62	0	991.04	4	1	TRUE
62	57	135.87	4	3	TRUE
62	114	120.06	4	6	TRUE
62	169	186.26	4	7	TRUE
62	224	499.19	4	9	TRUE
63	0	0.96	1	1	FALSE

Table 2.5 **Tumour volumes for stage IV melanoma cohort (selected indicative rows out of 34 rows)**
This table shows the CT imaging data for stage IV melanoma patients. Patient, number of patient in the study; days, days since start of treatment; volume_mL, tumour volume in mL; n_lesions, number of lesions in the patient; TP, timepoint the sample was taken; evaluable, denotes evaluability of lesion through CT imaging.

Characteristic	Stage IV melanoma	Stage II-III melanoma
Median age in years (range)	66 (38-66)	58 (23-79)
Gender - male (%)	6 (67%)	17 (45%)
Gender - female (%)	3 (33%)	21 (55%)
Total patients	9	38

Table 2.6 **Patient baseline characteristics**
Baseline characteristics for each of the patient cohorts in this study are shown.

Chapter 3

Detection of early-stage non-small cell lung cancers using personalised ctDNA analysis

3.1 Attribution

This chapter will be written up as a manuscript after handing in the thesis.

3.1.1 Author contributions

Wet lab work

I have carried out all the INVAR related lab work for this chapter, which also included organising the initial shipment of tissue, plasma and buffy coat samples from Papworth and Addenbrookes hospital to our lab. I extracted DNA from a total of 125 tissue samples, each comprising of 10 tissue slides which I macrodissected before extraction. After DNA extraction I performed DNA repair on samples with sufficient material. I then sonicated all tissue samples before preparing sequencing libraries from them. Upon successful library preparation I submitted the majority of tissue libraries for shallow whole genome sequencing (data not shown here) and captured all tissue libraries with exome bait sets. In parallel, I performed buffy coat extraction on a subset of the cohort (n=20) and organised for the remaining samples (n=80) to be extracted by an external collaborator. I sheared the 90 buffy coat samples of patients who had tissue available and prepared libraries followed by exome capture. I submitted all tumour and buffy coat captures for whole exome sequencing to

identify patient specific mutation lists which I then used to design custom capture panels for the plasma samples.

I organised the extraction of the plasma samples (n=90, coming from 360 aliquots) through an external collaborator. Upon return of extracted samples I quantified the DNA with dPCR before preparing libraries for these plasma samples. All samples were captured with the customised bait sets I designed based on the tumour and buffy coat sequencing.

For the collaboration with Inivata I facilitated the communication with the company and prepared the required information for the project. I also identified the plasma samples required for analysis that had to be sent to Inivata and organised their shipment.

Dry lab work

With help from Dineika Chandrananda and James Morries I performed mutation calling on the tissue and buffy coat whole exome sequencing data. I devised additional filters to remove remaining artefactual mutations from the original calling set before using the final mutation lists for the design of custom capture bait sets. Upon return of custom capture sequencing data from the plasma samples I aligned and collapsed the sequencing data before running the INVAR pipeline (chapter 2). For the subcohort of samples that was sent to Inivata I analysed the data returned to us further and compared it to the data I had generated with INVAR.

Writing of the chapter

I wrote the first draft of this chapter and generated all the figures presented here. I received comments and edits from Jonathan Wan, Irena Hudecova and Mareike Thompson to improve the chapter.

3.1.2 Funding

I would like to acknowledge the support of The University of Cambridge and Cancer Research UK (grant numbers A11906, A20240, C2195/A8466, and C9545/A29580). The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.337905.

3.2 Aims

The primary objective of this chapter was to apply the tool developed in chapter 2 to an independent cohort to assess its generalisability. In this chapter I investigate:

1. Whether INVAR can be applied to an independent cohort without major tweaking of individual parameters.

2. The detection rates in this cohort of samples with low levels of ctDNA and whether the new method proves advantageous to currently published methods.

3.3 Abstract

Although liquid biopsies have shown sensitive detection of cancer patients with early-stage disease across a variety of cancer types, detection in lung cancer has proven difficult and was only assessed in small cohorts. In this paper, we assess ctDNA levels in a large cohort of 100 patients with early-stage non-small cell lung cancer (NSCLC) prior to treatment with curative intent. The cohort contained 60 patients with stage I disease, allowing us to focus on patients where ctDNA detection has been challenging previously. Using custom capture sequencing panels, followed by an Integration of Variant Reads approach, we detected ctDNA in 62% of all patients, including 47% detection of patients with stage I disease and 81% and 89% detection for patients with stage II and III disease. ctDNA mutant allele fractions varied from $1.7\text{x}10^{-5}$ to $6\text{x}10^{-2}$ across this cohort, highlighting both the low levels of ctDNA and the variability in ctDNA fraction in this cancer type. Additionally, detection was superior in squamous cell carcinoma patients compared to adenocarcinoma patients (75% vs. 53% detection rate, respectively). To validate our results, we assessed the concordance between patient-specific mutation analysis and a generic sequencing panel, identifying concordance in 71% of cases (14 out of 17). Overall, the patient-specific approach was more sensitive on this subset of patients (detection of 82% vs. 65% for the untargeted approach). Our study characterises ctDNA in early-stage NSCLC patients using both patient specific and non-patient specific analysis platforms. Despite using a highly sensitive patient-specific analysis, still over 50% of early-stage cancers remain undetected, highlighting that ctDNA-based approaches on its own may not be sufficient in this setting and that using a multi-analyte analysis may provide a way forward.

3.4 Introduction

Despite the numerous advances in the field of liquid biopsy over the past decade [8], liquid biopsies have had limited success in circulating tumour DNA (ctDNA) detection in early-stage lung cancer. In the published literature, detection rates vary from 37% to 45% in stage I disease [19, 64, 98] with the exception of one study showing a 100% detection rate for a small cohort of stage IB non-small cell lung cancer patients (NSCLC) [36]. These studies use ctDNA to assess patients based on either small panels with patient-specific mutation information [19] or larger panels utilising cancer-specific mutation information based on public data sets [36, 64, 98]. In case of the CancerSEEK assay, additional tumour markers were used to further enhance cancer detection [98].

It is not clear why detection rates for ctDNA are low in early-stage NSCLC. It has been suggested that this might be due to low disease burden in early-stage, leading to limited release of ctDNA in the circulation which is not detected given the current sensitivity limitations of analytical platforms [19, 31]. Alternatively, it is possible that the reduced detection rates are due to differences in tumour biology in the early stages, whereby only more aggressive tumour forms are releasing sufficient amounts of DNA to be detected in the plasma [28].

In this study we assess the detection of ctDNA in 100 early-stage, treatment-naïve NSCLC patients. Where tumour tissue was available (n=90), we utilised patient-specific sequencing panels in combination with our previously published INVAR analysis pipeline (Chapter 2). INVAR utilises large, patient-specific mutation lists in combination with noise reduction methods, signal-weighting and signal integration in order to detect ctDNA. We also applied a commercial targeted sequencing assay [123] to samples from 27 patients who underwent chemo-/radiotherapy treatment, including those patients without tissue (n = 10, Fig. 3.1A). For both of these methods, ctDNA detection rates were analysed with regards to cancer stage and histological subtype.

3.5 Main text

A total of 100 treatment naïve patients with stage I – IIIB NSCLC undergoing surgery or chemo-/radiotherapy were recruited to the LUCID study (REC 14/WM/1072). The cohort predominantly consisted of stage IA and IB patients (n=60), recruiting only 19 patients with stage III disease (Fig. 3.1B). The median age of the cohort was 72 years (range: 44–88 years) and 89 (89%) of the recruited patients were current or previous smokers. Tumour subtype information was available for 82 patients and consisted of 34% (28 out of 82)

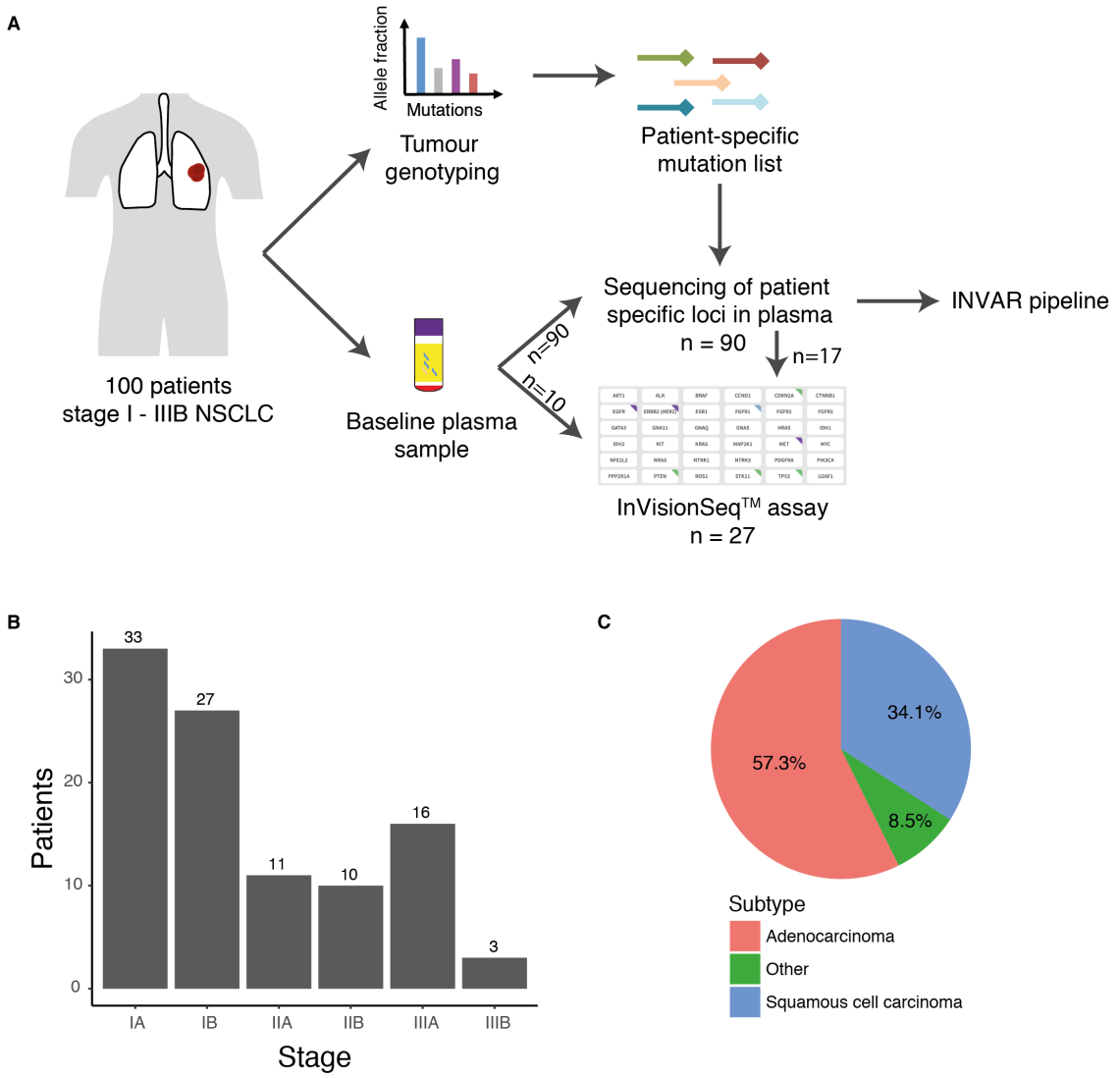


Fig. 3.1 **LUCID study design.**
(**A**) Patients were recruited to the LUCID study. Where tumour tissue was available, patient specific sequencing panels were designed and applied to baseline plasma samples (n=90). Custom capture sequencing data was analysed using the INVAR pipeline. Where tumour tissue was unavailable, patients were analysed using the InVisonSeq™ assay (n=10). 17 samples were analysed with both platforms. (**B**) Stage distribution in the LUCID cohort. 60% of the patients presented with stage I disease. (**C**) Subtype distribution in the LUCID cohort.

squamous cell carcinoma, 57% (47 out of 82) adenocarcinoma and 9% (7 out of 82) other subtypes (Fig. 3.1C). Patients either underwent surgery (71 out of 100) or radiotherapy ± chemotherapy (29 out of 100, supplementary table 3.2). For 90 patients, we obtained a tissue sample (either from surgery or remaining material from the diagnostic biopsy) and used it to guide our analysis. Plasma samples were obtained before the initiation of treatment, at each treatment visit (for patients undergoing radiotherapy ± chemotherapy), and during a nine-month follow up window.

For our tumour-guided analysis of this cohort we used the plasma samples that were collected at baseline before the initiation of treatment, and analysed them using our previously published INVAR pipeline (Chapter 2). Additionally, we applied the InVisionSeq™ assay [123] to cases where no tumour tissue was available and a subset of patients analysed with INVAR.

3.5.1 Mutational landscape in NSCLC

For patients with available tissue we analysed the abundance of commonly mutated lung cancer genes and mutation signatures. We identified commonly mutated lung cancer genes from the literature [88, 124] and compared their abundance to the present cohort. The results are shown in the heatmap in Fig. 3.2. For the two main subtypes of NSCLC in this cohort (adenocarcinoma and squamous cell carcinoma) TP53 is the most commonly mutated gene (46%), which agrees with the literature [88, 124]. Compared to the literature we see a surprisingly large number of KRAS mutations (32%). This can partially be explained by the subtype distribution in this cohort, which is biased towards adenocarcinomas (table 3.2) [88]. Mutation rates found in NF1 (23%), EGFR (16%), CDKN2A (12%) and BRAF (11%) are mostly comparable to those indicated in the literature (Fig. 3.2) [88, 124]. As the landscape of the two main tumour subtypes is quite different, it will be interesting to associate the observed tumour mutations with the respective subtypes once more detailed patient information becomes available. For example, KRAS and EGFR seemed more indicative of adenocarcinoma while CDKN2A is more commonly seen in squamous cell carcinoma [88, 124].

We also analysed the mutation signatures of the present cohort [125]. Using the deconstructSigs package [126] and the 26 signatures of mutation processes identified by Alexandrov and colleagues in 2013 [125], we computed the mutation signatures present in the 90 patients with available tissue data (Fig. 3.3). Signature 1A was most abundant in this cohort (82%). As described by Alexandrov and others, signature 1 is seen across all cancer types and is thought to be associated with ageing [125]. Given that the patients in this study were diagnosed at a median age of 72 (table 3.1), the high abundance of signature 1 is not surprising. The

second most frequently observed mutation signature is signature 4 (76%, Fig. 3.3), which is associated with the smoking of tobacco [125]. In this lung cancer cohort 90% of patients are current or previous smokers (table 3.1), which is a common observation in lung cancer cohorts and explains the abundance of signature 4. Apart from these two common signatures, the remaining signatures are observed in only few patients in this cohort and have not been previously associated with lung cancer [125].

3.5.2 Detection using the patient-specific INVAR pipeline

Whole exome sequencing information from the tumour and matched buffy coat was available for 90 of the 100 patients and yielded a median of 345.5 mutations per patient (IQR 213 to 515), which were used to design a total of three patient-specific hybrid-capture panels to cover all 90 patients. Sequencing panels were applied to baseline plasma samples from these 90 patients, and INVAR analysis was performed (Chapter 2). In this cohort, 99.8% of mutations were private, allowing patient samples to serve as controls for one another (Chapter 2). Signal was error suppressed on a cohort and sample basis and additionally weighted by both fragment length and corresponding tumour allelic fraction (Fig. 3.4A and B, Chapter 2). As tumour allelic fraction always also depends on tumour purity, one could think about adding an additional tumour purity feature to the INVAR pipeline or rely on cancer cell fractions instead of tumour allele fractions. For the present cohort this information was only available for just over 50% of samples and could, therefore, not be implemented. For samples with available information, a median of 20% estimated tumour purity was observed (IQR 0.1 - 0.3).

Detection limits were determined using receiver operating characteristic (ROC) curves and a sensitivity threshold of at least 95% was set (median=96.15%) on each of the three sequencing panels (Fig. 3.5A), detecting 62.2% (56 out of 90) of patients (see Supplementary table 3.7). Applying the same detection thresholds to sets of healthy controls that were run on the same panels, we observed a mean specificity of 95.84% (Fig. 3.5A). The patients with detectable levels of ctDNA had signal in a median of 2.6% of the initially targeted patient specific mutations (IQR 0.7% - 24.7%). As the majority of identified mutations were private to each patient, this minimal signal would have likely been missed with a generic panel or hotspot assay. Using the INVAR pipeline, we generated a median of 80,460 informative sequencing reads (IR) at patient-specific loci (IQR 42,030- 149,000). The number of IR guide in the estimation of the sensitivity limit in each sample (roughly 1/IR). In this cohort, 4 samples yielded less than 20,000 IR, limiting our ability to detect ctDNA below levels of $5x10^{-5}$ (Fig. 3.5B). Additionally, the INVAR pipeline can infer the total cancer genomes in the circulation at the time of blood draw by dividing the total number of unique

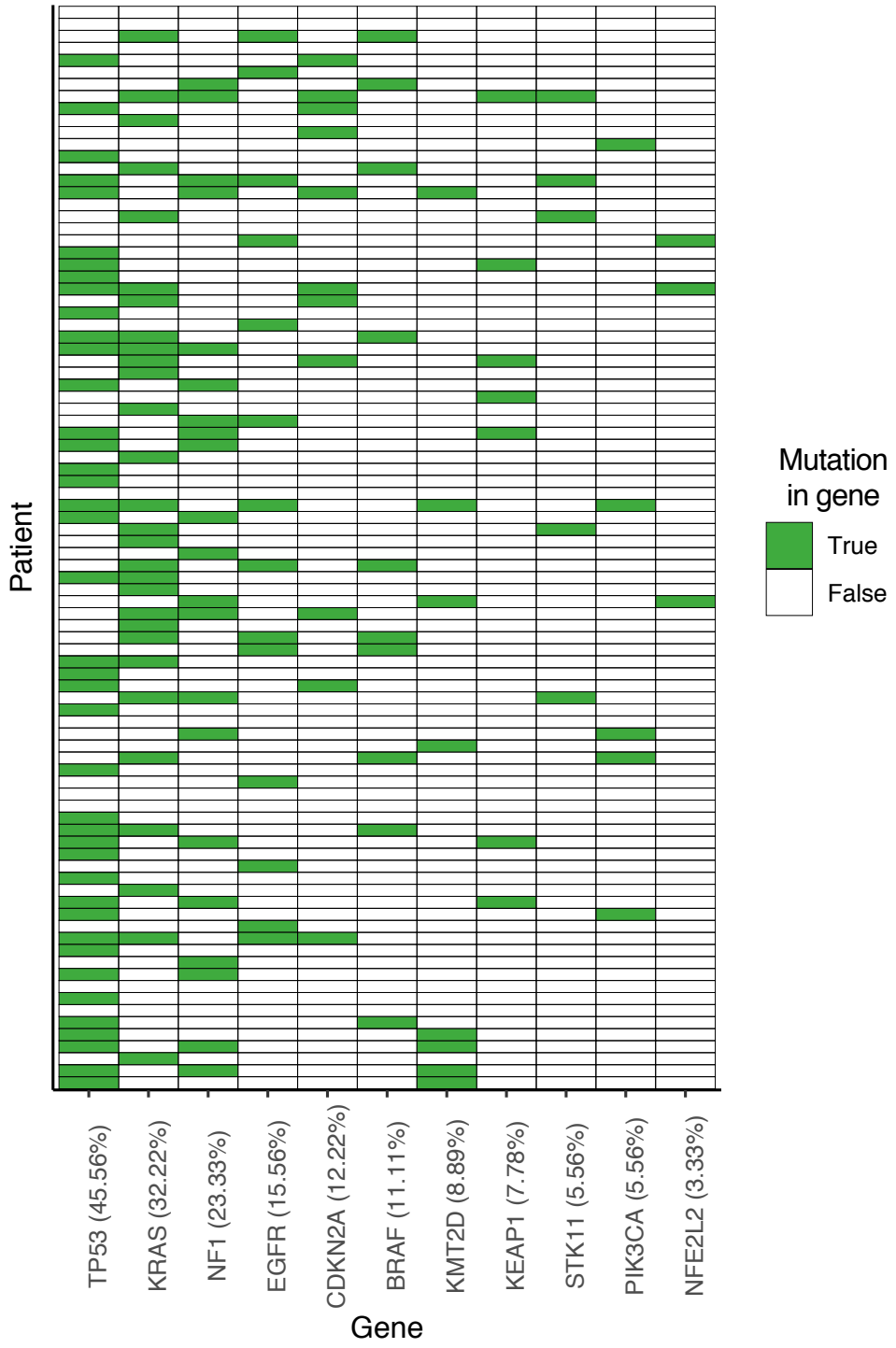


Fig. 3.2 **Commonly mutated lung cancer genes.**
Commonly mutated lung cancer genes were identified from Collisson et al. and Hammerman et al. [88, 124]. Abundance of mutations in these genes was determined in the present cohort where tumour tissue was available (n=90 patients). Boxes are highlighted in green for the presence of mutations in a given gene and patient.

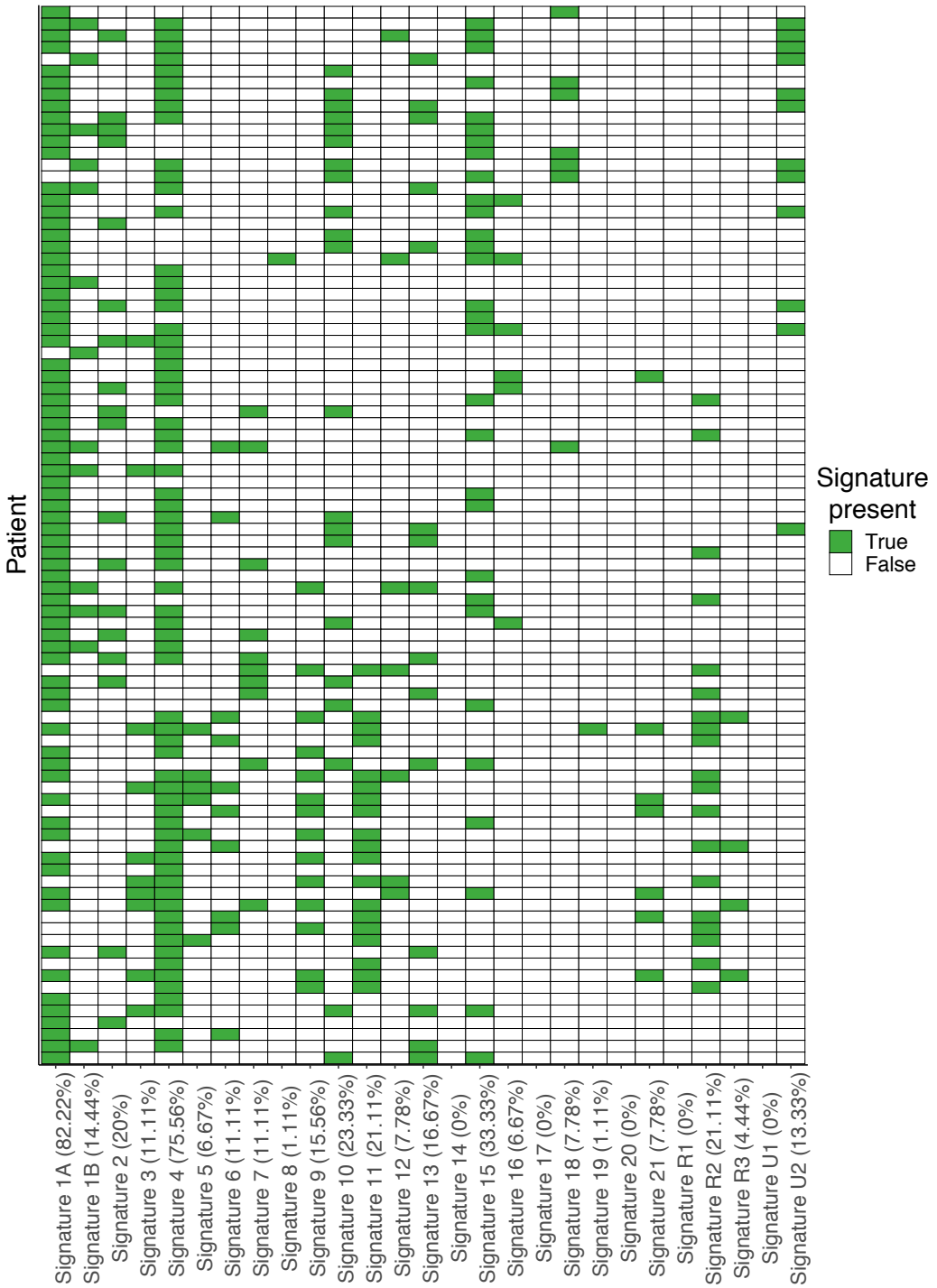


Fig. 3.3 **Mutation signatures in lung cancer.**
Mutation signatures were obtained using deconstructSigs [126] and are shown for the 90 patients in this study where tissue data was available. Boxes are highlighted in green for signatures present in the mutation data of a given patient.

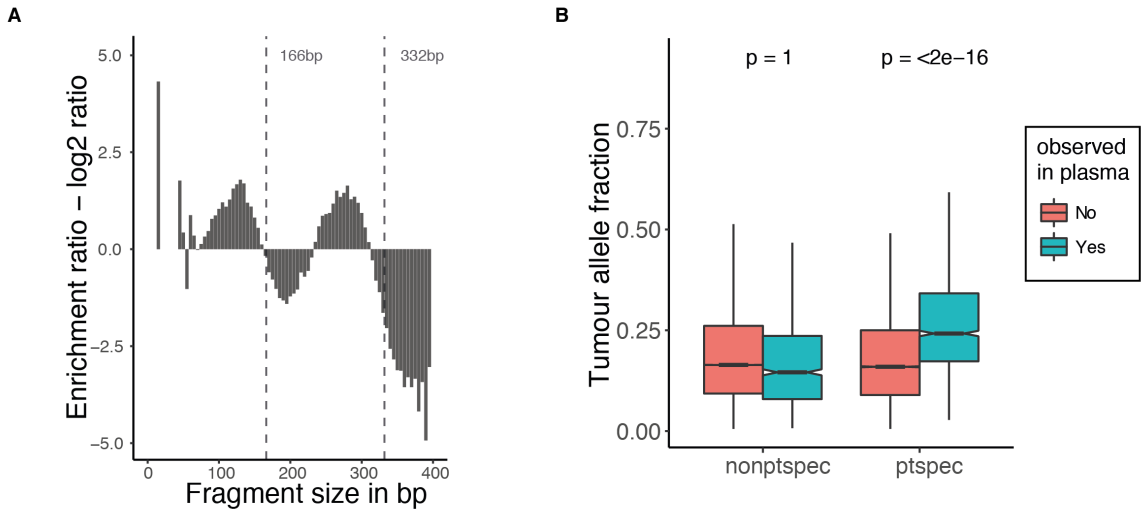


Fig. 3.4 **INVAR feature analysis in NSCLC.**
(**A**) Fragment length enrichment ratios in cohort. Enrichment ratios were used to weight mutant fragments by their fragment length. (**B**) Comparison of tumour allelic fractions of observed and not observed plasma mutations. In patient specific samples, mutations seen in the plasma come from tumour mutations with a higher allelic fraction. This was not observed in control samples (Wilcox-test, $p<2x10^{-16}$ vs 1).

mutated molecules with the total unique number of cfDNA molecules (Chapter 2). For the 56 samples with detected ctDNA from INVAR we observe a median cancer genome fraction of 0.074 (IQR 0.016 to 0.99) (Fig. 3.5C), highlighting the low levels of disease burden in the circulation in our early-stage cohort.

3.5.3 Detection of ctDNA using the InVisionSeq™ assay

Next, we applied the Inivata InVisionSeq™ assay [123] to 27 of the 100 patients, which were part of the chemo-/radiotherapy cohort. Ten of these had no tumour tissue available, rendering them ineligible for a patient-specific approach. The remaining 17 patients were chosen to compare ctDNA detection between INVAR and the InVisionSeq™ assay. The InVisionSeq™ assay analyses single nucleotide variants, copy number variants and insertions and deletions in regions from 36 selected cancer related genes, covering 10.61kb in total [127]. Based on the original TAm-Seq method [35], this enhanced TAm-Seq (eTAm-Seq™) technology uses 72bp-154bp long amplicons in a two-step multiplex PCR amplification to prepare sequencing libraries [127]. In a validation study across two laboratories Inivata could show 94% sensitivity and 99.99% specificity for mutations with allelic fractions ranging from 0.25% to 0.33% and could detect AFs as low as 0.02%[127].

A

	Sp. pt ctrl	Sp. healthy ctrl
Batch1	98.01%	97.40%
Batch2	95.42%	93.84%
Batch3	95.00%	96.29%
Mean	96.14%	95.84%

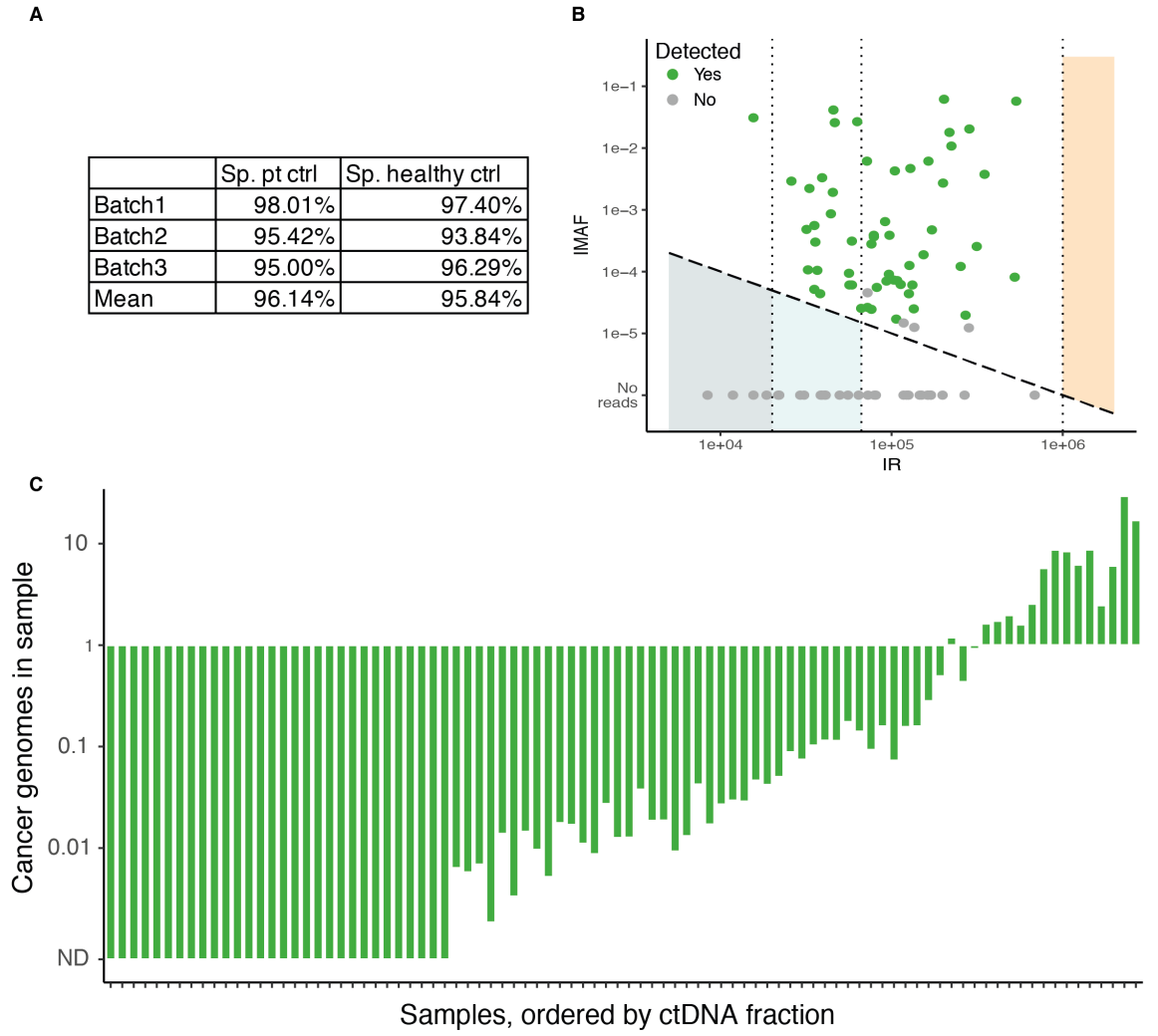


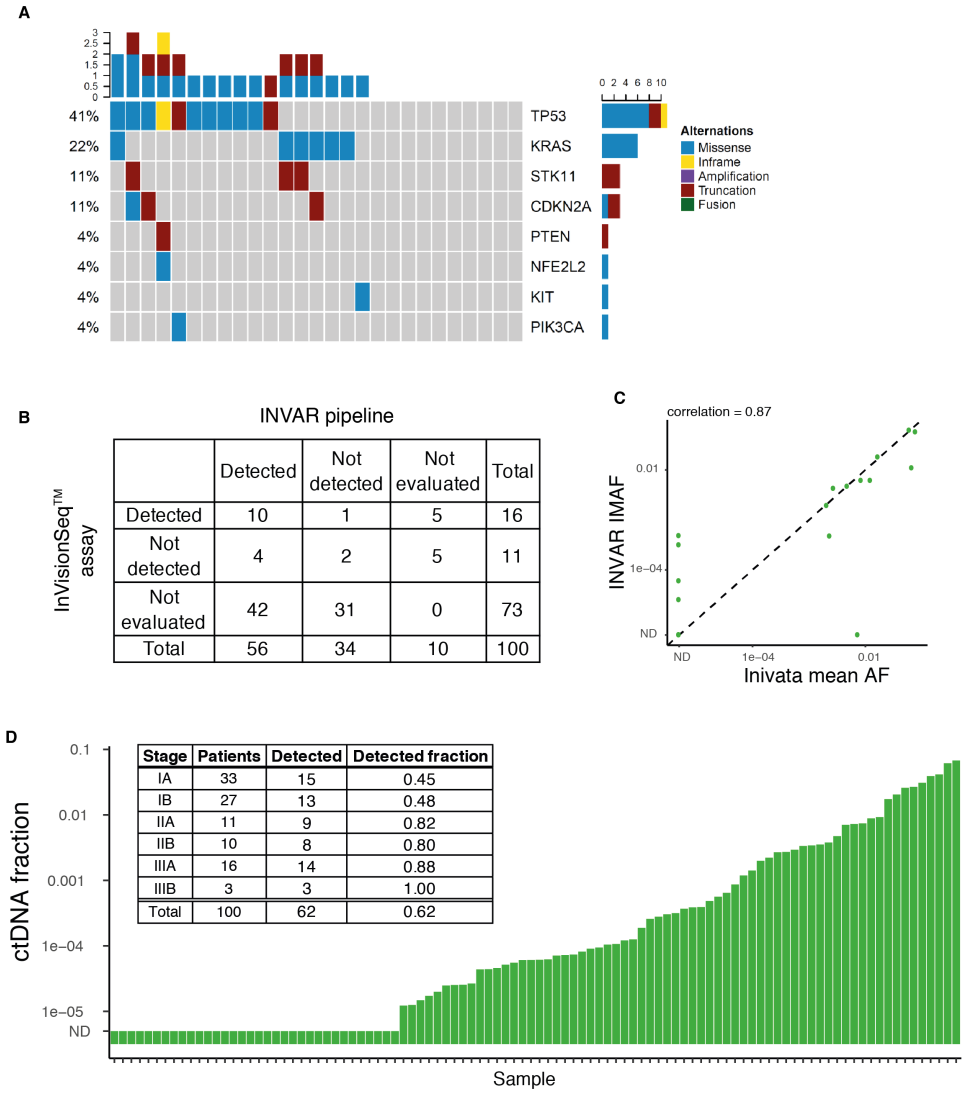
Fig. 3.5 **INVAR pipeline analysis in NSCLC.**
(**A**) Specificities for the three custom capture sequencing panels in this study. Specificities were assessed twice. Once using the other patients as controls and once using an independent set of healthy controls. (**B**) Informative Reads (IR) are plotted against Integrated Mutant Allele Fraction (IMAF). Sensitivity of the INVAR pipeline increases with increasing IR and can be roughly estimated as 1/IR (indicated by dashed diagonal line). Samples with less than 20,000IR (dark grey box) allow only for limited detection while samples with more than 1,000,000 IR (indicated by orange box) would allow for highly sensitive detection of ctDNA. Samples that pass the sensitivity threshold are shown in green. (**C**) Cancer genomes in the circulation. The total proportion of cancer genomes in the circulation was estimated for patients with detected levels of ctDNA.

Using a median input of 3.7mL of plasma (3.2mL – 4mL) and a median of 7,240 DNA copies (2,720 – 16,000), this non-patient-specific approach initially detected ctDNA in 17 of 27 patients in the present study. For 12 of these 17 patients matching tumour and buffy coat data was available. In one case this led to the reclassification of a mutation to a SNP (signal observed in the buffy coat), yielding a false positive rate of 8.3%. No tumour tissue was available for the other five detected patients. Overall, the InVisionSeq™ assay detected ctDNA in 16 of 27 patients (59%) with a mean of 1.67 alterations per patient and a total of 27 mutations across the cohort (Supplementary tables 3.3, 3.4, 3.5, and 3.7). 18 of these calls (66.7%) were below an allelic fraction of 0.01 and 12 (44.4%) were below an allelic fraction of 0.005, highlighting again the low levels of ctDNA in the early-stage setting. The majority of patients had alterations in TP53 (37%), followed by alterations in KRAS (22%), STK11 (11%), CDKN2A (11%), PTEN (4%), NFE2L2 (4%), KIT (4%) and PIK3CA (4%) (Fig. 3.6A). The performance of the InVisionSeq™ assay was best for stage III patients, reaching a detection of 73% (11 out of 15) (Supplementary tables 3.3, 3.4, 3.5). Detection decreases to 50% (3 out of 6) in stage II patients and 33% (2 out of 6) in stage I patients.

17 out of the 27 patients analysed with the InVisionSeq™ assay were also analysed with the INVAR pipeline as they had tumour tissue available and qualified for a patient specific analysis approach. 10 of the 17 cases (59%) were detected and two cases (12%) were not detected with both platforms, leading to a concordance between the two platforms of 71% (Fig. 3.6B). When comparing the ctDNA fractions of the detected cases between the two methods we observe a correlation of 0.87 (Pearson's r, p = 0.001, Fig. 3.6C). Out of the remaining discordant samples (n=5), four were only detected with the patient-specific INVAR pipeline but missed with the untargeted InVisionSeq™ assay (Fig. 3.6B). Interestingly, there was one case detected with the InVisionSeq™ assay but missed by the INVAR pipeline (Fig. 3.6B). This case had only 43 patient-specific mutations identified by whole exome sequencing, limiting the potential sensitivity of INVAR which was developed to target hundreds to thousands of mutations, thereby maximising the overall sensitivity in a given sample. This highlights that if only few patient specific mutations are available, a patient specific amplicon-based or even untargeted analysis approach might be preferred based on assay cost and time constraints.

3.5.4 Overall detection rates in NSCLC

62 of the 100 patients were detected positive at baseline (62%). When split up by stage, we observe detection rates of 47% for stage I, 81% for stage II and 89% for stage III patients (Fig. 3.6D). Similar to previous reports on ctDNA detection in NSCLC, we see the highest detection rates for the advanced stage patients and observe similar or higher detection rates



	Detected	Not detected	Not evaluated	Total
Detected	10	1	5	16
Not detected	4	2	5	11
Not evaluated	42	31	0	73
Total	56	34	10	100

Stage	Patients	Detected	Detected fraction
IA	33	15	0.45
IB	27	13	0.48
IIA	11	9	0.82
IIB	10	8	0.80
IIIA	16	14	0.88
IIIB	3	3	1.00
Total	100	62	0.62

Fig. 3.6 **InVisionSeq™ assay analysis in NSCLC.**
(**A**) A total of 27 mutations were identified; most alterations were identified in TP53 (37% of samples, after removing the false positive call), followed by KRAS (22%). Matched tissue data confirmed 70% (14/20) of these calls. (**B**) Comparison between INVAR and InVisionSeq™. Sensitivity of INVAR alone was 62% (56/90) and InVisionSeq™ detected ctDNA in 59% (16/27). 17 samples were analysed with both platforms with a concordance in ctDNA detection of 71% (12/17, 10 samples detected and two samples undetected with both platforms). (**C**) ctDNA correlation between INVAR and InVisionSeq™. 10 samples were detected with both platforms; yielding a correlation in ctDNA AF of 0.87 (Pearson's r, p = 0.001). (**D**) In total, ctDNA was detected in 62 of 100 samples. Detection rates are reported by stage in the inserted table.

compared to other reports [19, 36, 64, 98]. In this cohort, we quantified ctDNA down to a mutant allele fraction of $1.7x10^{-5}$. Overall, the median detected allelic fraction across all stages was $3.9x10^{-4}$ (IQR $7.2x10^{-5}$ to $3.7x10^{-3}$, Fig. 3.6D, Supplementary table 3.7).

Histological subtype information was available for 82 of the 100 patients. Similar to a previous report [19], we also observed higher detection rates in squamous cell carcinoma than in adenocarcinoma, with detection of 75% (21 out of 28), and 53% (25 out of 47), respectively. We also detected 71% (5 out of 7) of patients with other subtypes (Supplementary table 3.6). We observed a lower mean ctDNA fraction in adenocarcinoma patients (0.003) compared to squamous cell carcinoma patients (0.006) and patients with other subtypes (0.011, Supplementary table 3.6).

3.6 Discussion

Characterisation of the present cohort of early stage NSCLC patients revealed the presence of previously identified commonly mutated genes and signatures for lung cancer. TP53, KRAS and NF1 were the most commonly mutated genes in this cohort, concordant with previous findings [88, 124]. Similar to previously published literature we observe an abundance of the ageing related signature 1 and smoking related signature 4 in this cohort [125]. Once unblinded to patient demographics, these can be correlated with patient age and smoking status. For some of the patients in this cohort samples from multiple tumour regions were available. This could allow for analysis of tumour clonality, which in turn could be developed as an additional feature of signal weighting for the INVAR piepline. Unfortunately, no information on the location of the regions and their respective distances to each other were available for this cohort, prohibiting such an analysis.

We and others have shown that the use of patient-specific mutation assays can improve sensitivity and yield improved ctDNA detection [19, 35, 36]. In this study, we have applied both, patient-specific and targeted mutation assays to members of a cohort of 100 stage I – IIIB NSCLC patients. The INVAR pipeline, utilising large lists of patient-specific mutation lists combined with signal weighting and integration to detect ctDNA, detected 62.2% (56 out of 90) of patients, reaching over 85% (34 out of 40) detection levels for patients with stage II disease and above. For patients with stage I disease INVAR detected 45% of patients. Part of this cohort was analysed using the targeted InVisionSeq™ assay where detection levels by stage were lower. When comparing the two platforms, 23% (4 out of 17) of patients were only detected by INVAR and one patient was only detected with the InVisionSeq™. Taken together, the two platforms would have detected 88% (15 out of 17) of all patients, highlighting the power of multiple testing approaches to improve overall detection.

However, there are still cases being missed, requiring an even more sensitive platform. INVAR could achieve greater sensitivity by targeting more patient specific mutations (for example identified by whole genome sequencing of the tumour) or analysing larger plasma volumes. Recent reports have indicated the use of either a combination of multiple tumour markers or patient-specific platforms to increase the overall detection in a cohort compared to targeted or untargeted mutation assay approaches [8, 128, 129]. One example of combining multiple markers is the CancerSEEK assay. Using just the ctDNA assay, detection was at 22% but was increased to 59% (more than 2.5-fold increase) when considering other blood-based tumour markers [98]. Especially for patients with stage I disease, detection rates changed 10-fold from 4.3% to 43% when adding additional markers [98]. The CancerSEEK assay is one example highlighting how combining multiple cancer markers can substantially increase the chances of cancer detection and might set the standard for future ctDNA based assays.

3.7 Methods

3.7.1 Patient cohort

Samples were collected from patients enrolled on the LUCID (REC 14/WM/1072) study, a prospective and observational study on patients with stage I – IIIB non-small cell lung cancer (NSCLC) undergoing treatment (either surgery or radio- +/- chemotherapy) with curative intent. The primary endpoint of the study investigates the detection rates and levels of ctDNA at baseline in early-stage NSCLC. Patients were consented by a research/specialist nurse or clinician. The study was coordinated by the Cambridge Cancer Trials Unit-Cancer Theme, which also prospectively collected patient demographics and clinical outcomes. An overview of all patient demographics is given in supplementary tables 3.1 and 3.2.

3.7.2 Sample collection and processing

Tissue samples were obtained either from surgical specimens or diagnostic biopsies and processed as FFPE tissue. The tissue was sectioned in 8 µm sections with one slide set aside for H&E staining to guide tumour extraction. Plasma samples were collected before treatment initiation (at baseline). Additional plasma samples were collected to be analysed in the future. These included samples during treatment (for chemo-/radiotherapy cohort), after surgery (for surgical cohort), in three-month intervals for nine months following the completion of treatment and, if possible, at relapse. For all plasma time-points, peripheral blood was collected in S-Monovette 9mL EDTA tubes. Within an hour of blood draw, samples were centrifuged at 1,600g for 10 minutes before undergoing another centrifugation at 20,000g

for 10 minutes. Plasma was then stored at -80°C. A buffy coat sample was collected at the beginning of treatment.

3.7.3 Sample extraction

For FFPE samples, the stained H&E slide was used to identify regions of high tumour cellularity, which were then macro dissected from the other tissue slides. Samples were extracted using the QIAmp FFPE Tissue Kit (Qiagen) according to manufacturer's instructions with the following modifications: DNA was incubated at 56°C and 500rpm over-night and elution was carried out by applying 20 µL to the membrane twice. FFPE repair was carried out for samples containing more than 800 ng of DNA using the NEBNext® FFPE DNA Repair Mix (New England Biolabs) according to the manufacturer's instructions.

Buffy coat samples were extracted from up to 1 mL either manually or using the QIAsymphony platform (Qiagen). Samples were eluted in 70 µL (manual extraction) or up to 200 µL (QIAsymphony). For plasma extraction, 2-4 mL of the sample were extracted using the QIAsymphony platform with the QIAmp protocol. 24 samples were extracted in each batch, including a positive and negative control to monitor extraction efficiency. FFPE and genomic DNA were quantified using a dsDNA broad range assay on the Qubit fluorimeter (ThermoFisher Scientific). Plasma samples were quantified using a digital PCR with 55 cycles on a Biomark HD (Fluidigm) with a Taq-man probe against a 65bp region of the RPP30 gene (Sigma Aldrich) [35].

3.7.4 Library preparation

Using the Covaris LE220 (Covaris) according to manufacturer's instructions, tumour and buffy coat DNA were first sheared to a fragment length of 200 bp.15 µL volumes and the 8 microTUBE-15 AFA Beads Strip V2 were used and fragmentation patterns were randomly validated using a Bioanalyser (Agilent). A total of 100 ng (tumour) and 50 ng (buffy coat) of sheared DNA were used for library preparation with the ThruPLEX DNA-seq kit (Rubicon). The number of library cycles was adjusted to the sample input based on the manufacturer's recommendations.

Up to 15 ng of plasma DNA were used for library preparation with the ThruPLEX Tag-seq (Rubicon) or SureSelect XTHS kit (Agilent). Depending on the input the number of amplification cycles was varied according to the recommendations from the manufacturers. After all library preparations, qPCR (NEBNext® Library Quant Kit for Illumina® in the ROX low setting, New England Biolabs) and Bioanalyser or TapeStation (both Agilent) were used to determine library concentration and size.

3.7.5 Exome capture of tumour and buffy coat samples

The Illumina TruSeq kit with a 45 Mbp bait set (Illumina) was used for the exome capture of tumour and buffy coat samples after library preparation. Keeping tumour and buffy coat DNA separate, 250 ng of each library were multiplexed in three-plex reactions. To ensure compatibility with the ThruPLEX libraries, 1µL of i5 and i7 TruSeq HT xGen universal blocking oligos (IDT) were added at the hybridisation step and the volume of CT3 buffer was adjusted to 51 µL. Samples underwent two rounds of hybridisation, each lasting for 24 hours. After exome capture, sample QC was performed as described above and sequencing was carried out using a HiSeq4000 (Illumina).

3.7.6 Mutation calling in tumour tissue

Tumour mutation calling was carried out in two batches. For the first batch of FFPE tumour biopsies, mutation calling was performed with Mutect2 with the default settings: –cosmic v77/cosmic.vcf and –dbsnp v147/dbsnp.vcf. To maximise the number of mutations retained, all variants achieving Mutect2 pass were retained. Mutation calls were then filtered as follows:

1. Buffy coat mutant allele fraction equals zero
2. Mutation not in homologous region
3. Mutation not at a multiallelic locus
4. 1000 Genomes ALL and EUR frequency equals zero
5. A minimum unique tumour depth of 5.

For the second batch, mutations were identified using MuTect2, VarDict and Freebayes. The same filters as described above were applied to this cohort. In addition to retaining mutations passing MuTect2, mutations identified by both Freebayes and VarDict were also retained, thereby increasing our total number of targetable mutations. In some cases, multiple tumour regions were available for the same patient so mutation calling was carried out on the individual tumour regions as well as from the merged bam files of both regions.

3.7.7 Design of hybrid-custom capture panel and plasma capture

Following mutation calling, custom capture panels were designed using the SureDesign software (Agilent). Mutation lists of 19 to 35 patients were combined per capture panel and

a 1x tiling density and balanced boosting were used. 120 bp RNA baits were used for the panels, which varied in size from 2.138Mbp to 2.987Mbp. Mutation lists for each panel are shown in Supplementary table 3.8.

Plasma libraries were captured in singleplex or two-plex using 1000 ng of total input. SureSelect XT and SureSelect XTHS captures were carried out according to the manufacturer's instructions. For ThruPLEX Tag-Seq libraries, i5 and i7 blocking oligos (IDT) were added based on the manufacturer's recommendations [130]. Captures underwent 13 cycles of post-capture amplification and underwent the same QC as described above. Samples were sequenced on the HiSeq4000 platform (Illumina).

3.7.8 Read collapsing on plasma sequencing data

Known 5' and 3' adapter sequences were specified in a separate FASTA file and removed using Cutadapt v1.9.1. Using BWA-mem v0.7.13 with a seed length of 19, the trimmed FASTQ files were aligned to the UCSC hg19 genome. For ThruPLEX Tag-Seq libraries, read collapsing was carried out using CONNOR [67] with a consensus threshold of 90% (-f flag of 0.9) and a minimum family size of 2 (-s flag of 2). SureSelect XTHS libraries underwent read collapsing using the Agilent Genomics NextGen Toolkit (AGeNT) with a minimum family size of 2 [69].

3.7.9 Plasma analysis using INVAR pipeline

Plasma custom-capture sequencing data was analysed using the INVAR pipeline, which was developed previously (Chapter 2).

3.7.10 Plasma analysis using the InVisionSeq™ assay

Up to 4mL of plasma were sent to Inivata, where samples were extracted and analysed using the InVisionSeq™ assay [123]. The essay utilises the eTAm-Seq™ technology, which is based on the previously developed TAm-Seq method[35]. Sequencing libraries are prepared using a two-step PCR amplification with amplicons (72bp-154bp in length) covering 10.61kb across 36 cancer related genes. Upon analysis, data was returned to us for further exploration.

3.8 Supplementary tables

	Variable: n/N(%)	**Total**
Age	Total patients	100
	Mean (Standard deviation)	71 (8)
	Median (Minimum, Maximum)	72 (44, 88)
	Inter-quartile range	66, 76
Age group	<50	2/100(2)
	50-59 years	8/100(8)
	60-69 years	31/100(31)
	70-74 years	20/100(20)
	75 plus years	39/100(39)
Sex	Female	49/100(49)
	Male	51/100(51)
Smoking status	Never smoked	10/99(10.1)
	Ex-smoker	67/99(67.7)
	Current smoker	22/99(22.2)
Number of years smoked (for ex-and-current smokers)	Number	88
	Mean (Standard deviation)	38 (16)
	Median (Minimum, Maximum)	40 (1, 71)
	Inter-quartile range	25, 50
Number of years quit smoking (ex-smokers only)	Number	67
	Mean (Standard deviation)	18 (15)
	Median (Minimum, Maximum)	14 (0, 51)
	Inter-quartile range	3, 32

Table 3.1 **Summary of demographic variables and smoking history.** n-Number of patients; N-Total number of patients; %-Percentage of patients; SD-Standard deviation; Min-Minimum; Max-Maximum.

	Variable	**n/N (%)**
Disease stage at diagnosis	IA	33/100(33)
	IB	27/100(27)
	IIA	11/100(11)
	IIB	10/100(10)
	IIIA	16/100(16)
	IIIB	3/100(3)
Treatment plan for patients	Surgical	71/100(71)
	Non-surgical	29/100(29)
Tumour subtype	Squamous cell carcinoma	28/82(34.1)
	Adenocarcinoma	47/82(57.3)
	Large cell carcinoma	0
	Other	7/82(8.5)

Table 3.2 **Summary of radiological/pathological history.** n-Number of patients; N-Total number of patients; %-Percentage of patients.

	Variable	**Number n/N**	**ctDNA detection rate (95% CI)**[†]
Disease stage at diagnosis	IA	15/33	0.45 (0.28 to 0.64)
	IB	13/27	0.48 (0.29 to 0.68)
	IIA	9/11	0.82 (0.48 to 0.98)
	IIB	8/10	0.80 (0.44 to 0.97)
	IIIA	14/16	0.88 (0.62 to 0.98)
	IIIB	3/3	1.00 (0.29 to 1.00)
Tumour subtype	Squamous cell carcinoma	21/28	0.75 (0.55 to 0.89)
	Adenocarcinoma	25/47	0.53 (0.38 to 0.68)
	Large cell carcinoma	0	0
	Other	5/7	0.71 (0.29 to 0.96)
Summary	Overall	62/100	0.62 (0.52 to 0.72)

Table 3.3 **ctDNA detection summary full LUCID cohort.** CI – Confidence Interval; [†] - Clopper-Pearson confidence interval.

Statistics	IA	IB	IIA	IIB	IIIA	IIIB
Number	33	27	11	10	16	3
Mean (SD)	0 (0.001)	0.002 (0.008)	0.003 (0.006)	0.002 (0.003)	0.018 (0.022)	0.005 (0.003)
Median (Min, Max)	0 (0, 0.003)	0 (0, 0.042)	0.001 (0, 0.02)	0 (0, 0.009)	0.008 (0, 0.068)	0.005 (0.002, 0.007)
IQR	0, 0	0, 0	0, 0.003	0, 0.002	0, 0.029	0.002, 0.007

Table 3.4 **Summary of ctDNA level by disease stage at diagnosis.** SD – Standard deviation; Min – Minimum, Max – Maximum; IQR - Inter-quartile range

Statistics	Squamous cell carcinoma	Adenocarcinoma	Large cell carcinoma	Other
Number	28	47	-	7
Mean (SD)	0.006 (0.011)	0.003 (0.012)	-	0.011 (0.022)
Median (Min, Max)	0 (0, 0.042)	0 (0, 0.068)	-	0.002 (0, 0.061)
IQR	0, 0.004	0, 0	-	0, 0.009

Table 3.5 **Summary of ctDNA level by tumour subtype.** SD – Standard deviation; Min – Minimum, Max – Maximum; IQR - Inter-quartile range

	Variable	**Number n/N**	**ctDNA detection rate (95% CI)**[†]
Disease stage at diagnosis	IA	1/3	0.33 (0.01 to 0.91)
	IB	1/3	0.33 (0.01 to 0.91)
	IIA	2/4	0.50 (0.07 to 0.93)
	IIB	1/2	0.50 (0.01 to 0.99)
	IIIA	8/12	0.67 (0.35 to 0.90)
	IIIB	3/3	1.00 (0.29 to 1.00)
Tumour subtype	Squamous cell carcinoma	3/6	0.50 (0.12 to 0.88)
	Adenocarcinoma	5/10	0.50 (0.19 to 0.81)
	Large cell carcinoma	0	0
	Other	4/4	1.00 (0.40 to 1.00)
Summary	Overall	16/27	0.59 (0.39 to 0.78)

Table 3.6 **ctDNA detection summary LUCID sub-cohort analysed with the InVision-Seq™ assay.** CI – Confidence Interval; [†] - Clopper-Pearson confidence interval.

Patient	detection	AF	DP	IMAF	specificity	detection_INVAR	AF_inivata	detection_Inivata
1001	FALSE	0	NA	NA	NA	Not analysed	0	No
1002	TRUE	3.10E-02	15649	3.10E-02	1	Yes	NA	Not analysed
1004	TRUE	7.26E-03	71807	6.14E-03	1	Yes	8.38E-03	Yes
1005	TRUE	2.55E-05	66684	2.55E-05	0.958	Yes	0	No
1006	TRUE	2.68E-02	62989	2.68E-02	1	Yes	NA	Not analysed
1007	FALSE	0	NA	NA	NA	Not analysed	0	No
1008	TRUE	6.08E-05	56792	6.08E-05	0.958	Yes	0	No
1009	TRUE	6.14E-02	202377	6.23E-02	1	Yes	6.05E-02	Yes
1010	FALSE	0	122832	0	0	No	NA	Not analysed
1011	TRUE	2.94E-03	26001	2.94E-03	1	Yes	NA	Not analysed
1012	TRUE	7.37E-05	103235	7.37E-05	0.992	Yes	NA	Not analysed
1013	TRUE	2.05E-02	284812	2.05E-02	1	Yes	NA	Not analysed
1014	TRUE	3.35E-03	39240	3.35E-03	1	Yes	NA	Not analysed
1015	TRUE	2.23E-03	33137	2.23E-03	1	Yes	NA	Not analysed
1016	TRUE	1.72E-05	106499	1.72E-05	0.980	Yes	NA	Not analysed
1017	FALSE	0	15634	0	0	No	NA	Not analysed
1018	TRUE	4.74E-03	128821	4.71E-03	1	Yes	4.77E-03	Yes
1019	TRUE	7.25E-05	107875	7.25E-05	0.992	Yes	NA	Not analysed
1020	TRUE	2.59E-02	46540	2.59E-02	1	Yes	NA	Not analysed
1021	FALSE	0	NA	NA	NA	Not analysed	0	No
1022	FALSE	0	22039	0	0	No	NA	Not analysed
1023	TRUE	2.67E-05	72331	2.67E-05	0.984	Yes	NA	Not analysed
1024	TRUE	2.80E-04	76461	2.80E-04	1	Yes	NA	Not analysed

Table 3.7 continued from previous page

Patient	detection	comb_AF	DP	IMAF	specificity	detection_INVAR	AF_inivata	detection_Inivata
1025	TRUE	1.99E-03	45383	1.92E-03	1	Yes	2.06E-03	Yes
1026	FALSE	0	267378	0	0	No	NA	Not analysed
1027	TRUE	1.05E-04	36917	1.05E-04	0.994	Yes	NA	Not analysed
1028	FALSE	0	55559	0	0	No	NA	Not analysed
1029	FALSE	0	40793	0	0	No	NA	Not analysed
1030	TRUE	4.86E-04	31818	4.86E-04	0.997	Yes	0	No
1031	TRUE	3.78E-03	350494	3.78E-03	1	Yes	NA	Not analysed
1032	TRUE	2.48E-05	76529	2.48E-05	0.984	Yes	NA	Not analysed
1033	TRUE	8.68E-04	44153	8.68E-04	1	Yes	NA	Not analysed
1034	TRUE	6.45E-04	91864	6.45E-04	1	Yes	NA	Not analysed
1035	TRUE	7.42E-03	11862	0	0	No	7.42E-03	Yes
1036	TRUE	1.26E-04	127338	1.26E-04	0.997	Yes	NA	Not analysed
1037	TRUE	1.41E-03	172565	4.79E-04	1	Yes	2.35E-03	Yes
1038	FALSE	0	8418	0	0	No	NA	Not analysed
1039	TRUE	1.21E-04	253961	1.21E-04	1	Yes	NA	Not analysed
1040	TRUE	8.87E-03	NA	NA	NA	Not analysed	8.87E-03	Yes
1041	TRUE	7.05E-03	NA	NA	NA	Not analysed	7.05E-03	Yes
1042	FALSE	0	149917	0	0	No	NA	Not analysed
1043	TRUE	3.69E-04	79053	3.69E-04	1	Yes	NA	Not analysed
1044	TRUE	3.53E-03	104630	4.31E-03	1	Yes	2.75E-03	Yes
1045	FALSE	0	164264	0	0	No	0	No
1046	FALSE	0	146286	0	0	No	NA	Not analysed

Table 3.7 continued from previous page

Patient	detection	comb_AF	DP	IMAF	specificity	detection_INVAR	AF_inivata	detection_Inivata
1047	TRUE	1.74E-02	218349	1.81E-02	1	Yes	1.67E-02	Yes
1048	TRUE	3.42E-03	NA	NA	NA	Not analysed	3.42E-03	Yes
1049	TRUE	1.07E-04	32513	1.07E-04	0.976	Yes	NA	Not analysed
1050	FALSE	0	81254	0	0	No	NA	Not analysed
1051	TRUE	6.07E-05	58887	6.07E-05	0.977	Yes	NA	Not analysed
1052	TRUE	2.69E-03	NA	NA	NA	Not analysed	2.69E-03	Yes
1053	TRUE	3.16E-04	58630	3.16E-04	0.997	Yes	0	No
1054	FALSE	0	NA	NA	NA	Not analysed	0	No
1055	TRUE	3.91E-02	224117	1.09E-02	1	Yes	6.73E-02	Yes
1056	FALSE	0	NA	NA	NA	Not analysed	0	No
1057	TRUE	6.23E-05	113274	6.23E-05	0.987	Yes	NA	Not analysed
1059	TRUE	5.63E-04	35454	5.63E-04	0.998	Yes	NA	Not analysed
1060	FALSE	0	685328	0	0	No	NA	Not analysed
1061	TRUE	4.40E-05	38242	4.40E-05	0.954	Yes	NA	Not analysed
1062	TRUE	9.05E-05	96433	9.05E-05	0.990	Yes	NA	Not analysed
1063	TRUE	3.04E-04	35974	3.04E-04	0.995	Yes	NA	Not analysed
1064	FALSE	0	49710	0	0	No	NA	Not analysed
1065	FALSE	0	39283	0	0	No	NA	Not analysed
1066	FALSE	0	169528	0	0	No	NA	Not analysed
1067	TRUE	4.16E-02	45812	4.16E-02	1	Yes	NA	Not analysed
1068	TRUE	5.16E-05	35424	5.16E-05	0.962	Yes	NA	Not analysed
1069	TRUE	9.45E-05	56509	9.45E-05	0.983	Yes	NA	Not analysed

Table 3.7 continued from previous page

Patient	detection	comb_AF	DP	IMAF	specificity	detection_INVAR	AF_inivata	detection_Inivata
1070	TRUE	4.37E-05	126500	4.37E-05	0.984	Yes	NA	Not analysed
1071	FALSE	0	18676	0	0	No	NA	Not analysed
1072	TRUE	9.27E-03	164237	6.16E-03	1	Yes	1.24E-02	Yes
1073	TRUE	3.92E-04	97321	3.92E-04	1	Yes	NA	Not analysed
1074	FALSE	0	198670	0	0	No	NA	Not analysed
1075	FALSE	0	38635	0	0	No	NA	Not analysed
1076	TRUE	2.58E-04	315301	2.58E-04	1	Yes	NA	Not analysed
1077	FALSE	0	72732	0	0	No	NA	Not analysed
1078	TRUE	1.88E-04	153617	1.88E-04	1	Yes	NA	Not analysed
1079	TRUE	3.88E-04	78860	3.88E-04	0.999	Yes	NA	Not analysed
1080	FALSE	1.25E-05	136320	1.25E-05	0.918	No	NA	Not analysed
1081	TRUE	6.13E-05	132775	6.13E-05	0.982	Yes	NA	Not analysed
1082	TRUE	2.53E-05	134923	2.53E-05	0.957	Yes	NA	Not analysed
1083	TRUE	5.54E-05	82169	5.54E-05	0.953	Yes	NA	Not analysed
1084	TRUE	1.99E-05	270906	1.99E-05	0.962	Yes	NA	Not analysed
1085	FALSE	0	64313	0	0	No	NA	Not analysed
1086	TRUE	8.18E-05	524149	8.18E-05	1	Yes	NA	Not analysed
1087	FALSE	0	162134	0	0	No	NA	Not analysed
1088	FALSE	0	117555	0	0	No	NA	Not analysed
1089	FALSE	0	30838	0	0	No	NA	Not analysed
1090	FALSE	0	79660	0	0	No	NA	Not analysed
1091	FALSE	0	21804	0	0	No	NA	Not analysed

Table 3.7 continued from previous page

Patient	detection	comb_AF	DP	IMAF	specificity	detection_INVAR	AF_inivata	detection_Inivata
1092	FALSE	1.23E-05	282719	1.23E-05	0.946	No	0	No
1093	TRUE	7.11E-05	93367	7.11E-05	0.973	Yes	NA	Not analysed
1094	FALSE	0	125872	0	0	No	NA	Not analysed
1095	FALSE	1.48E-05	117972	1.48E-05	0.935	No	NA	Not analysed
1096	TRUE	1.19E-03	NA	NA	NA	Not analysed	1.19E-03	Yes
1097	TRUE	2.72E-03	199838	2.72E-03	1	Yes	NA	Not analysed
1098	TRUE	6.78E-02	536373	5.77E-02	1	Yes	7.79E-02	Yes
1099	FALSE	0	169632	0	0	No	NA	Not analysed
1100	FALSE	4.59E-05	72611	4.59E-05	0.943	No	NA	Not analysed
1101	FALSE	0	41324	0	0	No	NA	Not analysed
1102	FALSE	0	29307	0	0	No	NA	Not analysed

Table 3.7 **ctDNA detection in LUCID cohort**
Listed are all patients in the LUCID study with their ctDNA detection levels. Denoted are patient ID, as well as detection and comb_AF (across both platforms). Detailed are also specific outputs of the INVAR application: DP (depth), IMAF (Integrated Mutant Allele Fraction), specificity and detection. Also listed are outputs from the application of the InVisionSeq™ assay: AF_Inivata and detection_Inivata

Chr	Position	Ref	Alt	Gene	Depth	AF	Patient	Mut class
chr17	72341041	G	C	KIF19	140	0.07	1032	G/C
chr2	166747053	G	C	TTC21B	125	0.19	1022	G/C
chr3	100949948	A	G	IMPG2	138	0.20	1024	A/G
chr2	141665427	C	A	LRP1B	39	0.05	1023	C/A
chr7	140287448	T	C	DENND2A	57	0.32	1026	T/C
chr1	80916992	C	A	HNRNPA1P64	142	0.07	1023	C/A
chr2	179412560	A	T	TTN	107	0.38	1026	A/T
chr1	157737268	C	A	FCRL2	184	0.24	1012	C/A
chr2	24261216	T	C	C2orf44	118	0.10	1032	T/C
chr22	25053485	G	A	POM121L10P	64	0.25	1017	G/A
chr8	100821666	A	G	VPS13B	258	0.09	1026	A/G
chr2	189917652	C	G	COL5A2	58	0.40	1012	C/G
chr3	62306099	G	A	C3orf14	98	0.11	1026	G/A

Table 3.8 **LUCID patient-specific mutation lists (selected indicative rows out of 38853 rows)**
This table contains all patient-specific mutation lists for patients in the LUCID study. Mutation positions are given using the hg19 genome build. For each mutation the chromosome, position and reference and alternate allele are given. Additionally, the gene name, tumour depth and allelic fraction, patient name and mutation class are also listed.

Chapter 4

Detection of ctDNA from dried blood spots after DNA size selection

4.1 Attribution

This chapter is adapted from a manuscript, which was submitted to bioRxiv in September 2019:

"Detection of ctDNA from dried blood spots after DNA size selection"

Katrin Heider[*], Jonathan C. M. Wan[*],..., Nitzan Rosenfeld[§]

[*] Equally contributing authors

[§] Corresponding author

4.1.1 Author contributions

K.H., J.C.M.W. and N.R. wrote the manuscript. K.H. and J.C.M.W. carried out the experiments and data analysis. J.H. and S.B. generated and prepared the animal model. I.H. and W.N.C. helped in designing the study. P.G.C., N.R. and D.G. led and coordinated work on the MelResist study from which the human blood spot was used. C.G.S., J.D.B. and N.R. supervised the project.

Wet lab work

Jonathan and I jointly developed the concept for blood spot collection, extraction and size selection. The preparation of each blood spot sample requires four steps: the blood spot has to be generated, extracted, size selected with magnetic beads and a library has to be prepared before submission for sequencing. After the extraction and bead selection we would also

perform a bioanalyser to confirm the removal of contaminating genomic DNA. The initial experiments to optimise the protocol (n = 10 blood spots) were carried out together. The later experiments to generate the data presented here were carried out by myself (n = 2 blood spots). The plasma and tumour shallow whole genome sequencing data of the melanoma patient presented in this chapter were prepared by Jonathan as part of chapter 2. The data of the ascites and engrafted tumour of the xenograft sample were generated by the Brenton lab.

Dry lab work

I devised the analysis pipeline for both, the xenograft and human blood spot data. For the xenograft data I identified Xenomapper [131] as a suitable package to differentiate between human and mouse reads and wrote a wrapper around it. For both the human and xenograft data I identified ichorCNA [75] as a tool to visualise copy number changes as a means to report the presence of ctDNA and wrote a wrapper around the publicly available code. I also analysed the plasma and tissue data (generated by Jonathan and the Brenton lab) with ichorCNA. Furthermore, I analysed and plotted the fragment patterns for the human blood spot as well as the copy number similarities between blood spot and matched plasma data presented in this chapter.

Writing of the manuscript

The first draft of the manuscript was written by me and I prepared all the figures presented in this chapter. After an initial draft, Jonathan and I worked jointly on improving the manuscript and communicated with Nitzan for further guidance. I then converted the manuscript into the chapter presented here, altering the figures and text where needed.

4.1.2 Acknowledgements

The authors would like to thank Catherine Thorbinson, Emily Barker and Alex Azevedo from the MelResist study group and the Cambridge Cancer Trials Centre, Addenbrookes Hospital, Cambridge. We also thank Carolin Sauer for acquisition of mouse samples and preparation of sequencing data.

4.1.3 Competing interests

N.R., J.D.B. and D.G. are co-founders, shareholders and officers or consultants of Inivata Ltd, a cancer genomics company that commercialises ctDNA analysis. Inivata had no role in the conceptualisation, study design, data collection and analysis, decision to publish or

preparation of the manuscript. Cancer Research UK has filed patent applications protecting methods described in this manuscript.

4.1.4 Funding

We would like to acknowledge the support of The University of Cambridge and Cancer Research UK (grant numbers A11906, A20240, C2195/A8466, and C9545/A29580). The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.337905.

4.2 Aims

The primary objective of this chapter was to develop a method for easier collection and storage of blood samples to allow less well equipped centres to participate in research studies. Together with Jonathan Wan I established a method:

1. Which allows blood samples to be stored longer term without hampering the quality of cfDNA for downstream analysis.

2. Which can process whole blood samples by removing the genomic DNA proportion and retaining the cfDNA proportion of interest.

3. For the analysis of ctDNA from as little as a single drop of blood by analysing changes in copy number in both human and xenograft samples.

4. To analyse the similarities between blood spot and plasma or tumour based sequencing data.

4.3 Abstract

Recent advances in the research and clinical applications of circulating tumour DNA (ctDNA) is limited by practical considerations of sample collection. Whole genome sequencing (WGS) is increasingly used for analysis of ctDNA, identifying copy-number alterations, fragment size patterns, and other genomic features. We hypothesised that low-depth WGS data may be generated from minute amounts of cfDNA, and that fragment-size selection may be effective to remove contaminating genomic DNA (gDNA) from small volumes of blood. There are practical advantages to using dried blood spots as these are easier to collect, facilitate serial sampling, and support novel study designs. Novel designs include the utilisation of archival samples by the removal of gDNA in small volumes, prospective human studies and studies based on animal models. We therefore developed a protocol for the isolation and analysis of cell-free DNA from dried blood spots. Analysing a dried blood spot of 50 µL frozen whole blood from a patient with melanoma, we identified ctDNA based on tumour-specific somatic copy-number alterations, and found a fragment size profile similar to that observed in plasma DNA processed by traditional methods. We extended this approach to detect tumour-derived cell-free DNA in a dried blood spot from a mouse xenograft model and were able to identify ctDNA from the originally grafted ascites. Together, our data suggests that ctDNA can be detected and monitored in dried blood spots. This will enable new approaches for sample collection from patients and *in vivo* models.

4.4 Introduction

Circulating tumour DNA (ctDNA) can be used to sensitively detect and quantify disease burden using a variety of sequencing based approaches [8]. For example, using shallow whole genome sequencing (sWGS), ctDNA can be detected down to mutant allele fractions of $\sim$3% through analysis of somatic copy-number alterations (SCNAs)[75, 77]. Alternatively, leveraging differences in fragment size between tumour-derived and non-tumour cell-free DNA molecules (cfDNA) can enhance the detection of genomic alterations and the identification of plasma samples from patients with cancer compared to healthy individuals [27, 132]. Although sWGS generates data on only a fraction of a single genome (0.3 genome equivalents correspond to $\sim$1 pg DNA), sequencing libraries for sWGS have traditionally been generated from larger amounts of cfDNA extracted from millilitre volumes of plasma from a venous blood samples [77]. Established protocols for collection of plasma for ctDNA analysis require prompt spinning of EDTA-containing tubes or delayed spinning of tubes containing cell preservatives/fixatives [133]. However, this processing restriction poses practical limitations on possible clinical study designs, especially in the context of serial sampling.

The use of limited blood volumes and dried blood spots for analysis of cfDNA may facilitate new trial designs, widen clinical applications, and enable point-of-care testing and self-collection of samples. Additionally, analysis of minute amounts of blood may facilitate longitudinal ctDNA monitoring from animal models with limited circulating blood volume. In prenatal diagnostics, polymerase chain reaction (PCR) has been used to carry out fetal RHD genotyping and HIV detection using maternal dried blood spots [99, 100]. In applications to cancer, ctDNA from a limited plasma volume was previously analysed in a study of a mouse xenograft model, where quantitative PCR was used to measure the human long interspersed nuclear element-1 (hLINE-1) as a measure of tumour burden [102]. In another pilot study in breast cancer patients, whole genome amplification was performed on blood obtained from a finger prick. They found comparable allelic frequencies in somatic mutations between the finger prick sample and matched venous blood [101]. Sensitive detection of ctDNA from limited volumes or blood spots represents a technical challenge due to the limited total number of mutant molecules. Whole blood samples are considered inferior to carefully-collected plasma samples due to the presence of contaminating genomic DNA (gDNA) from lysed white cells in whole blood [8, 134], which dilutes tumour-derived ctDNA signal. In this study, we present methods for cfDNA extraction from dried blood spots and the subsequent analysis and detection of ctDNA.

4.5 Results

We sought to assess the number of cfDNA genome copies that can be sequenced from a single blood drop or dried blood spot. Based on previous reports, the median concentration of cfDNA is approximately 1600 amplifiable copies per mL of blood for patients with advanced cancer [64, 135]. This translates to approximately 80 copies of the genome as cfDNA in a blood drop/spot of 50 µL. Assuming a yield in the range of ∼60%-80% in DNA extraction and efficiency of ∼15%-40% in generating a sequencing library, this is estimated to result in approximately 7x-25x representation of the genome in sequencing libraries prepared from cfDNA from a single blood drop. We therefore hypothesised that low-depth WGS of cfDNA can be attainable from a dried blood spot after removal of genomic DNA.

To test this hypothesis, we thawed frozen whole blood from a patient with Stage IV melanoma, and transferred 50 µL to a Whatman FTA filter paper card. After drying the card for 15 minutes, we performed DNA extraction and library preparation from the dried blood spot. An overview of the workflow is shown in Fig. 4.1A. Quality control using capillary electrophoresis revealed contaminating gDNA, as indicated by an excess of large DNA fragments (Fig. 4.1B). cfDNA fragments typically display a characteristic fragmentation profile with a prominent peak at 166bp [15, 27]. This peak was not observed, likely due to the low mass of cfDNA in the blood spot and the larger amounts of gDNA. We attempted a library preparation from the extracted blood spot but were unsuccessful, likely due to the inability of the longer fragments to undergo successful amplification. To remove these contaminating gDNA fragments (> 500bp in length), we applied a right-side size-selection using AMPure beads, at a bead-to-sample ratio of 1:1 (Section 4.7). For this selection, beads are added to the sample and initially the DNA bound to the beads (carrying high molecular weight DNA) is removed while the supernatant containing low molecular DNA is retained. In a second step, additional AMPure beads are added (in a bead-to-sample ratio of 7:1) to capture all the remaining small-size fragments in the supernatant [136]. We generated a sequencing library from the size-selected DNA using the Thruplex Tag-Seq kit, and obtained a total of 232,107,928 sequencing reads (PE150; Illumina HiSeq4000; Fig. 4.1A).

In our data, we achieved a unique sequencing depth of 6x from sWGS following collapsing with a minimum family size of 1 [67]. Using a diversity estimator (SPECIES [137]), we inferred that, based on the distribution of the number of duplicate reads per molecule, up to 10x unique coverage are likely to be achieved from this blood spot library (Section 4.7).

Sequencing data obtained from the blood spot was analysed for somatic copy-number alterations using ichorCNA [75].The generated copy number plot is shown in Fig. 4.2A. The alterations observed were consistent with those identified in a matched plasma sample from the same patient, isolated by standard plasma DNA-based methods (Fig. 4.2A). The extent of

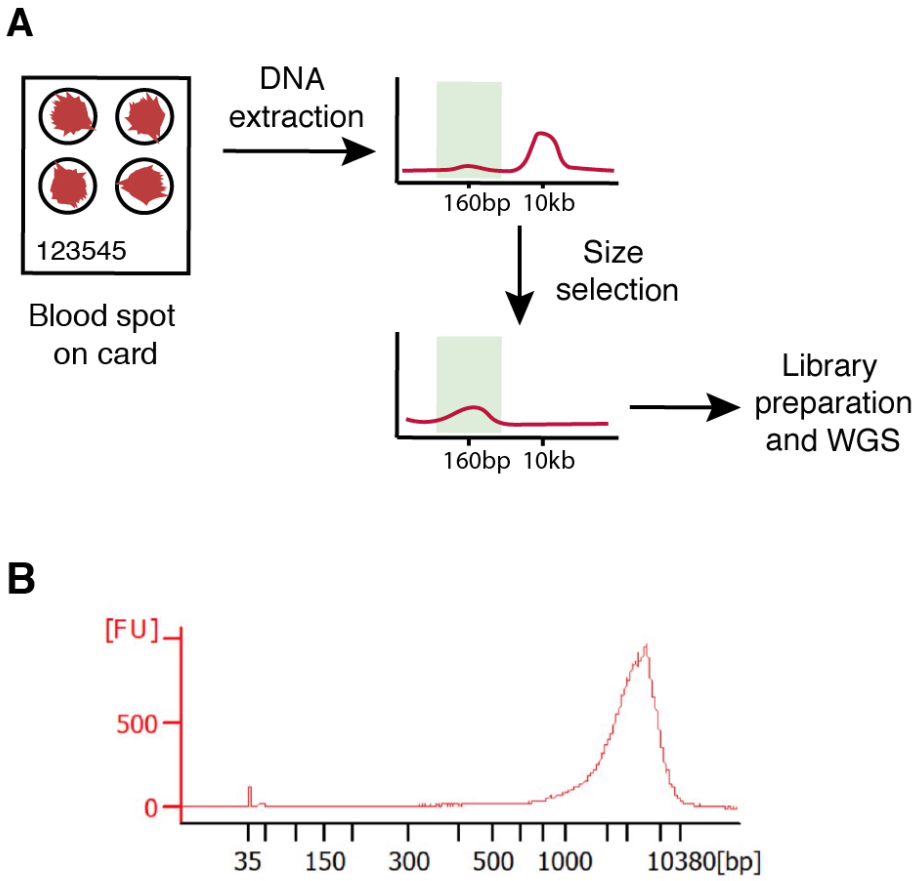


Fig. 4.1 **Schematic representation of experimental workflow.**
(**A**) Overview of the analysis of dried blood spots: DNA extraction, followed by size selection, and low-depth WGS. (**B**) Bioanalyser trace of DNA extracted from a 50 µL dried blood spot from blood of a patient with advanced melanoma, showing a high level of genomic DNA contamination (> 1kbp) and no clear cfDNA peak (∼166bp).

SCNAs between the two samples was significantly correlated (Pearson's r = 0.75, p < 2.2 x 10^{-16}, Fig. 4.2B) and similar to that found in the initial tumour biopsy copy number profile (Fig. 4.2A).

Using sWGS, we show that the overall fragment size distribution of the human blood spot cfDNA was comparable to that of cfDNA derived from plasma (Fig. 4.2C) [8, 14, 27]. We then independently analysed the size distribution of mutant and wild-type reads, leveraging mutation calls from exome sequencing of matched tumour tissue in order to accurately distinguish true mutations from sequencing noise. This confirmed that the tumour-derived fragments were shorter in size compared to wild-type fragments, with modal sizes of 150 bp and 170 bp, respectively (Fig. 4.2D). These data recapitulate size profiles derived from plasma samples of cancer patients [8, 14, 27].

We applied INVAR to the blood spot data and were able to detect ctDNA and could compare the obtained IMAF to that of the time-matched plasma. INVAR was applied to custom capture data of technical replicates of the plasma sample as well as WES and sWGS data. Informative reads (IR), the total count of observed mutant reads and IMAF of these samples are compared in table 4.1 below. As to be expected, the largest number of IR were observed for the technical replicates of the custom capture data where input material and sequencing saturation are the highest. We observe five times more IR in the blood spot sWGS data compared to the sWGS of the matched plasma, likely due to the difference in sequencing depth between the two data-sets. When comparing the obtained IMAF, a good concordance is observed between the different INVAR applications to the plasma sample. Interestingly, the blood spot sample has a lower IMAF, potentially indicating remaining contaminating DNA fragments diluting down the ctDNA.

sample type	**data type**	**Informative reads**	**mut sum**	**IMAF**
plasma	custom capture	940,770	72,359	0.070
plasma	custom capture	1,233,542	107,006	0.079
plasma	WES	124,920	8,889	0.060
plasma	sWGS	5,529	383	0.069
blood spot	sWGS	26,867	1,110	0.039

Table 4.1 **INVAR application to blood spot data.** INVAR was applied to the blood spot data of the melanoma patient and the time-matched plasma sample. Informative reads, total count of mutant reads (mut sum) and the obtained integrated mutant allele fraction (IMAF) are shown.

We next considered whether blood spot analysis may have applications in the longitudinal analysis of disease burden in live murine patient derived xenograft (PDX) models. At

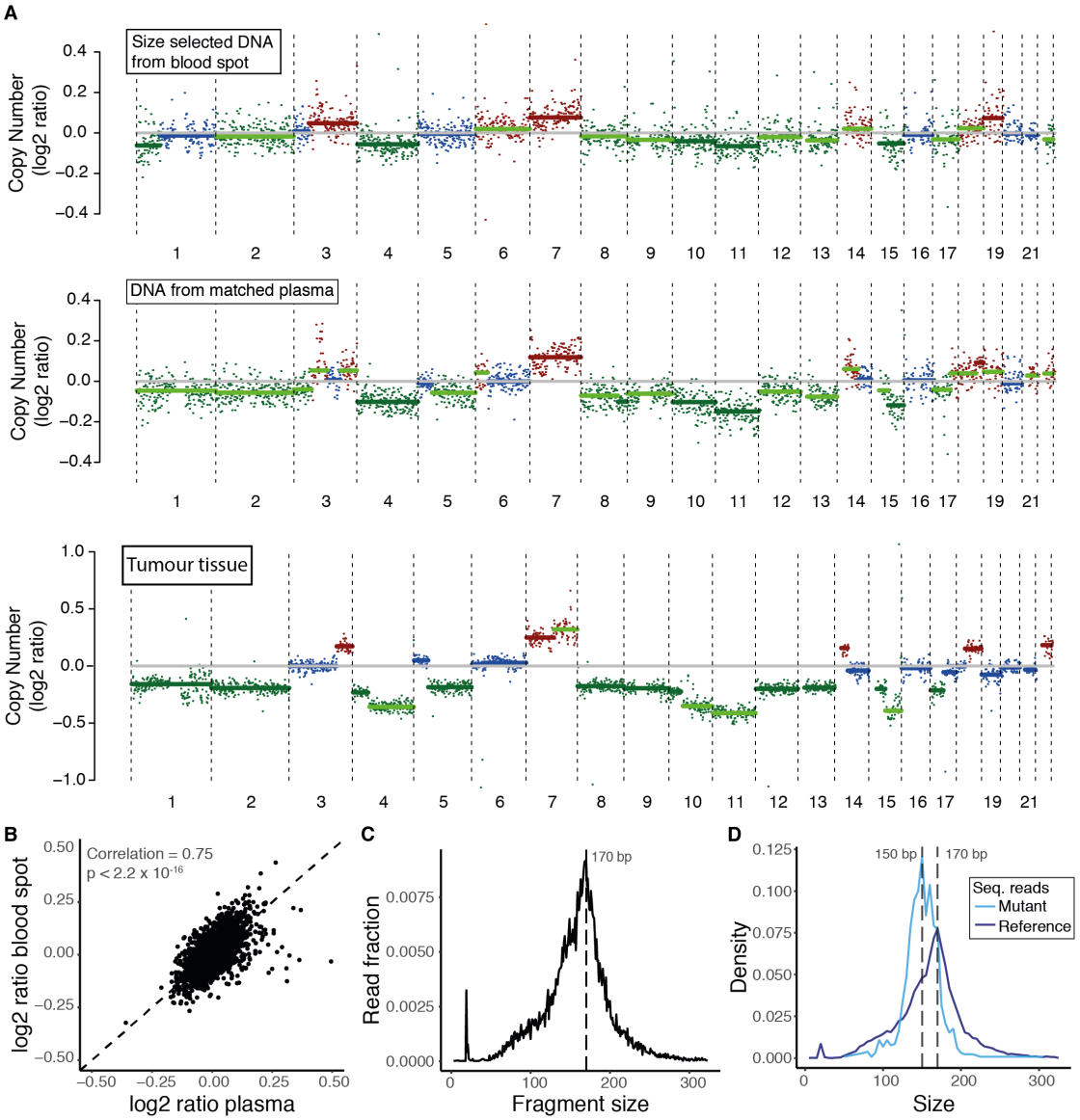


Fig. 4.2 **Detection of ctDNA in a dried blood spot from a cancer patient.**
(**A**) Copy-number profiles from sWGS of a sequencing library generated from the same dried blood spot as in Fig. 4.1B after size selection, from a matched plasma sample from the same individual and time point and the matched tumour tissue. Blue=neutral, red=gain, green=loss. (**B**) Correlation of log2 ratios for each copy-number bin using ichorCNA [75], comparing bins between matched blood spot and plasma data. The correlation in log2 ratios for all bins between the two samples was 0.75 (Pearson's r, $p < 2.2 \times 10^{-16}$). (**C**) Size profile of the sequencing reads generated from the size selected blood spot DNA library (data shown in panel A). The overall size profile is comparable to that of cfDNA, i.e. with a peak at ∼170 bp. (**D**) Length of the sequencing reads (data from panel C) carrying known patient-specific mutations (light blue), and reads carrying reference alleles at the same loci (dark blue).

present, analysis of cfDNA is challenging in small rodents as the volumes of blood required for most traditional ctDNA analysis can only be obtained through terminal bleeding. To assess the feasibility of dried blood spot analysis in animal models, we sampled 50 µL of whole blood onto a dried blood spot card from an orthotopically implanted ovarian tumour PDX model. DNA was extracted and sequenced (Section 4.7). Following alignment of sequencing reads, both human genome (tumour-derived) and mouse genome (wild-type) reads were observed, again showing characteristic fragmentation patterns of mutant and wild-type cfDNA (Fig. 4.3A) [27]. Copy-number alterations were observed when analysing the human sequencing reads and mirrored the profile observed in both the original patient ascites sample and the matched PDX tumour in the mouse (Fig. 4.3B). This confirms that blood spots can indeed be used to monitor disease progression and burden in animal models.

4.6 Discussion

In this study, we demonstrate a new method to detect ctDNA in blood drops/spots using sWGS from both human and PDX samples. Our analysis mirrors observations previously made in cfDNA plasma analysis. This approach relies on the use of size-selection to remove genomic DNA, combined with ctDNA measurement approaches such as sWGS which leverage signal from across the entire genome. Only highly multiplexed approaches leveraging signals from multiple loci are suitable for blood spot analysis. The analysis of any individual locus would have limited sensitivity due to the small number of genome copies of cfDNA that may be obtained from a single blood spot (in the order of 5-50 copies).

We analysed a dried blood spot from a patient with melanoma and observed a good correlation in the copy-number profiles obtained from the blood spot and a time-matched plasma and tumour sample. We see similar cfDNA and ctDNA size profiles as observed from standard plasma DNA-based methods. Further work on larger cohorts with fresh finger prick blood is warranted before progressing towards broader use of blood spots for ctDNA monitoring. This work should compare the ctDNA allele fractions between blood spot and matched plasma, as the extent of gDNA contamination of blood spot DNA might vary. Additionally, the sensitivity limit for ctDNA analysis in blood spots should be determined with both sWGS and targeted sequencing approaches. If single nucleotide variants were to be targeted, a larger number of patient-specific mutations must be identified and interrogated in order to adequately mitigate the effect of sampling error from the limited copies of cfDNA in small volumes of blood. In future, the potential application of personalised sequencing panels to sequencing data could facilitate highly sensitive monitoring of disease from even small volumes.

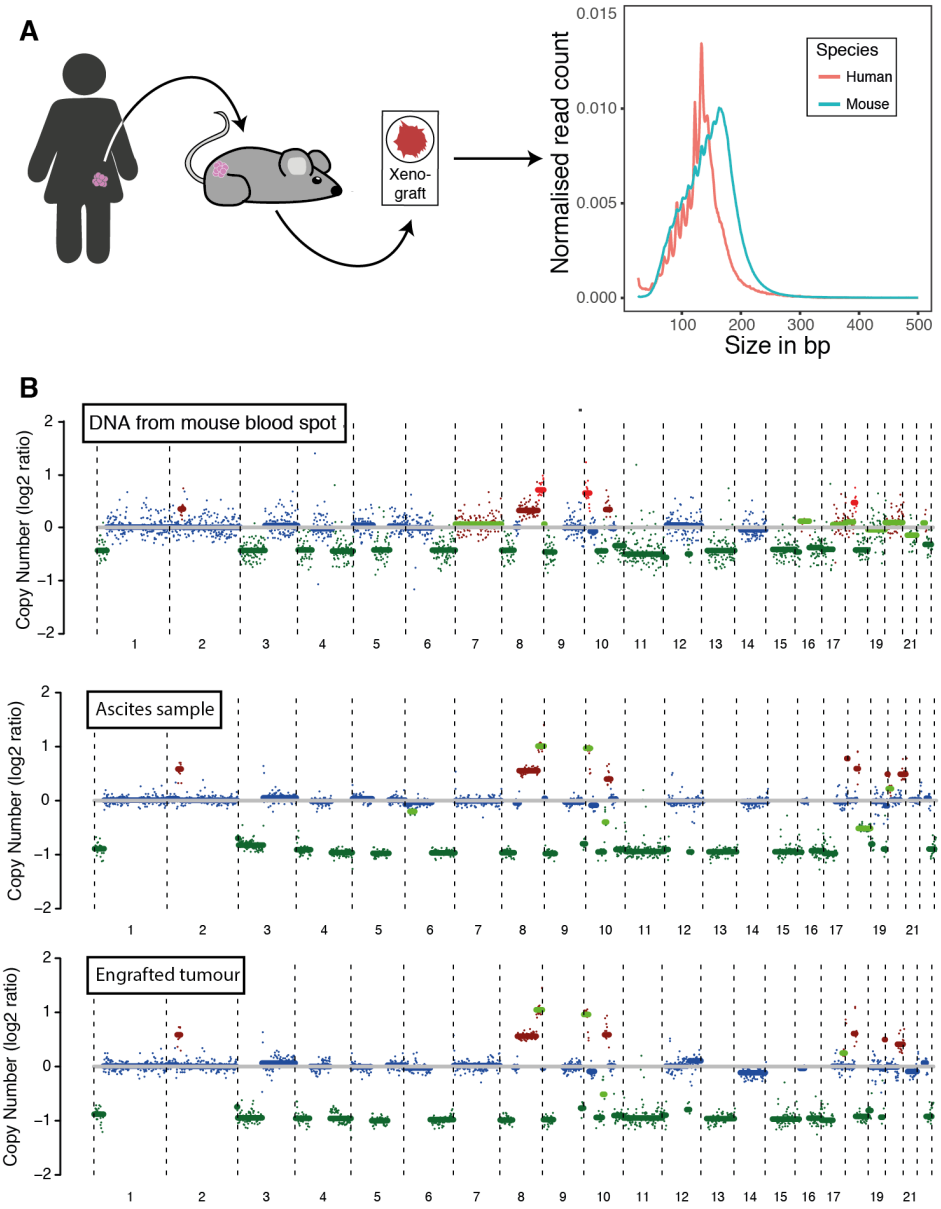


Fig. 4.3 **ctDNA detection from a dried blood spot in a xenograft model.**
(**A**) sWGS analysis of whole blood taken from a mouse xenograft model of ovarian cancer (illustrated in the left panel). The fragment lengths of reads aligning to the human genome (red, representing tumour ctDNA) were shorter than those aligning to the mouse genome (blue, representing non-tumour cfDNA). (**B**) Copy-number profiles were successfully generated from a dried blood spot from the mouse ovarian xenograft model (Section 4.7). The copy-number profiles of the original human ascites sample and the engrafted tumour are also shown. Segments coloured in blue, red and green indicate regions of copy-number neutrality, gain and loss, respectively.

In addition, we demonstrate the value of this approach in animal models, allowing the detection of SCNAs and the characteristic ctDNA fragmentation pattern from dried blood spots of PDX models. In the monitoring of ctDNA in small animal models, overcoming low circulating blood volumes is a major challenge. Although tail vein blood sampling in rodents has already been used for longitudinal cancer monitoring from small blood samples, analysis was limited to high copy-number markers such as hLINE repeat sequences [102]. Here, we highlight the possibility of next generation sequencing of blood spot cfDNA, enabling both, shallow and up to 10x WGS.

From a practical standpoint, the application of dried blood spots could enable high-frequency ctDNA monitoring of patients and animal models. Sampling and pre-analytical processing can be further simplified, potentially supporting new study designs incorporating wider populations and more frequent collection of smaller sample volumes. We hope that detection of ctDNA from limited blood volumes will enable novel approaches for cancer monitoring, such as self-collection of samples at home followed by shipping and centralised analysis.

4.7 Methods

4.7.1 Cell-free DNA extraction from dried blood spots

The human sample was collected from a patient enrolled on the MelResist study (REC 11/NE/0312). Written consent to enter the study was taken by a research/specialist nurse or clinician who was fully trained regarding the research. MelResist is a translational study of response and resistance mechanisms to systemic therapies of melanoma, including BRAF targeted therapy and immunotherapy, in patients with stage IV melanoma. Thawed whole blood (50 µL) was transferred to Whatman FTA™ Classic Cards (Merck) and allowed to air dry for 15 minutes before DNA extraction. A single fresh blood spot was obtained from an ovarian cancer xenograft mouse model immediately after culling, and similarly applied to Whatman FTA™ Classic Cards, and allowed to air dry. Blood spot card samples were stored at room temperature inside a re-sealable plastic bag. DNA was extracted from the card using the QIAamp DNA Investigator kit (Qiagen), using the manufacturer's recommended extraction protocol for FTA and Guthrie cards, which are conventionally used for assessment of inherited genetic conditions in neonates from gDNA. The protocol was followed with the following modifications. 1) three 3mm punches were made from the blood spot, and carrier RNA was added to Buffer AL as per the manufacturer's recommendation. 2) blood spot

DNA (which we hypothesised contained both cfDNA and gDNA) was eluted in two rounds of 25 µL elution buffer.

4.7.2 Size-selection and library preparation of blood spot cfDNA

Blood spot DNA eluates contain a low concentration of cfDNA, among a large background of gDNA (Fig. 4.1B). cfDNA library preparation cannot be effectively performed from such a sample since the abundance of long fragments reduces the likelihood of any cfDNA fragments successfully being ligated with adaptor molecules for subsequent amplification. Based on our characterisation of gDNA length of >1-10kb (Fig. 4.1B), and previous work demonstrating that cfDNA in vitro ranges from ∼70-300bp in length with a peak at 166bp [7], we opted to perform size-selection in order to remove contaminating long gDNA fragments. Thus, a right-side size-selection was performed on DNA eluates using AMPure XP beads (Beckman Coulter) in order to remove long gDNA fragments. For this purpose, we adapted a published protocol for a right-side size selection that is conventionally used for DNA library size-selection prior to next generation sequencing [136]. Following optimisation of bead:sample ratios for cfDNA fragment sizes, we used a bead:sample ratio of 1:1 to remove contaminating gDNA. The supernatant was retained as part of the right-side size-selection protocol. A second size-selection step used a 7:1 bead:sample ratio to capture all remaining fragments, and the size-selected DNA was eluted in 25 µL water. Blood spot eluates were concentrated to 10 µL volume using a vacuum concentrator (SpeedVac), since this volume is the maximum recommended for downstream library preparation using the Thruplex Tag-Seq kit (Takara). 16 cycles of library amplification were carried out. Libraries underwent QC using Bioanalyser 2100 (Agilent) and qPCR with the Illumina/ROX low Library Quantification kit (Roche) on a QuantStudio 6 (Life Technologies). Libraries were submitted for whole-genome sequencing on a HiSeq4000 (Illumina) with paired end 150bp/cycles.

4.7.3 Plasma library preparation

Plasma cfDNA libraries were prepared for the matched time point where the blood spot was collected as well as a cohort of 49 healthy controls. The DNA was extracted using the QIAsymphony (Qiagen) with the QIAamp protocol and quantified by digital PCR on a Biomark HD (Fluidigm) using a 65bp TaqMan assay for the housekeeping gene RPP30 (Sigma Aldrich) [35] and 55 cycles of amplification. Using the estimated number of RPP30 DNA copies per µL eluate, the cfDNA concentration in the original sample was estimated. Up to 9.9ng were used for the library preparation. The ThruPLEX Tag-Seq kit (Takara) was used according to the manufacturer's instructions and 7 cycles of amplification were

carried out. After barcoding and sample amplification, the library underwent bead clean-up and underwent QC as described above. The sample was submitted for sequencing on a HiSeq4000 with paired end 150bp/cycles.

4.7.4 Tumour library preparation

For the human blood spot, a time-matched tumour sample was available. Tumour DNA was extracted as described by Varela et al. [103] and sheared to 200bp fragment length using the COVARIS LE220 Focused-ultrasonicator according to manufacturer's instructions. 50ng of material were prepared for sWGS using the ThruPLEX Plasma-Seq kit (Takara) according to the manufacturer's instructions and 7 cycles of amplification were carried out. After barcoding and sample amplification, the library underwent bead clean-up and underwent QC as described above. The sample was submitted for sequencing on a HiSeq4000 with 150bp/cycles.

For the xenograft sample, material from the engrafted tumour as well as the human ascites sample used for grafting were available for analysis. The sample was extracted using the Qiagen allprep kit (Qiagen) and the DNA was sheared to 200bp fragment as described above. 50ng of DNA were prepared with the Thruplex DNA-Seq kit (Takara) according to the manufacturer's instructions and followed by a bead clean-up (1:1 ratio, as described above). The sample was quantified using TapeStation (Agilent) and submitted for sequencing on a HiSeq4000 with single end 50bp/cycles.

4.7.5 Sequencing data analysis

All samples were sequenced on a HiSeq4000. FASTQ files were aligned to the UCSC hg19 genome using BWA-mem v0.7.13 with a seed length of 19, then deduplicated with MarkDuplicates. For sWGS detection of ctDNA, ichorCNA was run as described [75], utilising a panel of normal derived from a set of plasma cfDNA samples from 49 healthy individuals (SeraLabs).

For xenograft sequencing analyses, BAM files underwent alignment to the mouse and human genomes in parallel using Xenomapper [131]. Fragment lengths were determined for both files using Picard CollectInsertSizeMetrics [138]. Additionally, ichorCNA [75] was run on the subset of reads aligning to the human genome to confirm the presence of CNA.

4.7.6 Library diversity estimation

In order to estimate the total number of cfDNA genome copies present in a blood spot library, we used CONNOR [67] to perform deduplication of the blood spot sequencing library based on endogenous barcodes [58] with minimum family sizes ranging between 1 and 5 (data not shown). For each family size setting, the mean deduplicated coverage was calculated using Samtools mPileup. Deduplicated coverage values for each family size setting were used as input for diversity estimation using a statistical method, SPECIES [137], best known for estimating the diversity of ecological populations based on the frequency of members observed through a random sample. A minimum family size of 1 was used for the data analysis.

Chapter 5

Discussion

The two main goals of my PhD were to improve the detection of ctDNA and to detect ctDNA from limited sample volumes or input copies of DNA. The development of the INVAR pipeline provides a means to detect ctDNA more sensitively down to levels of parts per million (chapter 2) and was applied to a large cohort of early-stage NSCLC patients to assess ctDNA levels (chapter 3). In chapter 4 I explored the possibility for a simpler method of liquid biopsy sampling by collecting blood spots, and showed ctDNA detection from minimal volumes of blood.

5.1 Improving the sensitivity of ctDNA detection

ctDNA is widely used in the cancer field and detection methods have evolved over time. Currently, the most sensitive methods in the ctDNA field focus on the detection of SNVs. Assays can be tumour guided, requiring *a priori* knowledge of tumour specific alterations. They may also focus on cancer (subtype) specific mutations (identified through public datasets). Finally, when ctDNA levels are sufficiently high, *de novo* mutation calling from the blood sample itself can be carried out by e.g. whole exome sequencing.

The most sensitive detection is achieved when tumour information is available and can help to detect low levels of ctDNA, which is why INVAR utilises patient specific tumour information for the detection of ctDNA (see chapter 2). Detection is improved further by a reduction in background error and an enrichment of ctDNA specific signal. Background errors are reduced by using UMIs and a sample specific suppression of outlier signal while signal enrichment is accomplished by utilising locus specific information on fragment length and tumour allelic fraction. Combined, INVAR reliably detected ctDNA down to parts per million when applied to a dilution series of plasma DNA obtained from a patient with melanoma. As INVAR requires prior tumour information, its application is limited to the

availability of a tumour sample. Therefore, INVAR is most suited to the detection of minimal residual disease after surgery when a tumour sample is readily available. The early detection of minimal residual disease can expedite treatment of disease relapse and improve the patient's prognosis. Similarly, cases with undetected MRD can provide a confidence of cure for the patient, reducing overtreatment.

In chapter 3, I applied INVAR to a cohort of early-stage NSCLC patients (from the LUCID study) to study the levels of ctDNA detectable in early-stage cancer. While ctDNA detection has yielded promising results in detecting cancer in patients with advanced disease and larger tumour burden, the detection rates in early-stage cancer have been low for many cancer types. By applying the INVAR assay to this cohort, I aimed to improve the detection of ctDNA and better understand the distribution of ctDNA levels in early-stage NSCLC. While the observed detection rates were similar to those observed in other smaller lung cancer studies (Section 1.4.1), the LUCID cohort contained patients at lower levels of ctDNA compared to the other studies [19, 36, 64, 98]. The demographic details with regards to stage and subtype distribution vary substantially between the different studies, making a direct comparison challenging. Nonetheless, the application of INVAR to the LUCID cohort provided additional information on the ctDNA levels that can reliably be detected in early-stage cancer.

For any assay the limit of sensitivity will depend on both assay specific limitations and sample specific restrictions. For example, the sensitivity of a SNV based assay will not only depend on its background error rate and the assay chemistry but also on the copies of DNA available for analysis and the number of mutations that can be targeted. The number of targetable mutations will depend on the cancer type: cancer types with a high mutation burden, such as melanoma and lung cancer, will allow for greater sensitivity compared to cancers that are more driven by copy number changes such as ovarian and breast cancer [84, 116]. As described in the introduction, these sample specific sensitivity considerations can be simplified to a two-dimensional matrix (Fig. 5.1).

The sensitivity can be improved by increasing the DNA material available for analysis or increasing the total number of targeted mutations. The DNA input can be increased by sampling a larger volume of blood or (depending on the assay) reducing sample loss by optimising the protocols. While PCR based methods usually retain the majority of the input molecules, library preparation based methods often lose >50% of the initial molecules. An improved library preparation approach that retains more molecules would immediately increase the total number of molecules, and therefore sensitivity, of a given assay. Simultaneously, sensitivity can be improved by increasing the number of targeted mutations. In the present INVAR application, mutation identification was based on prior whole exome

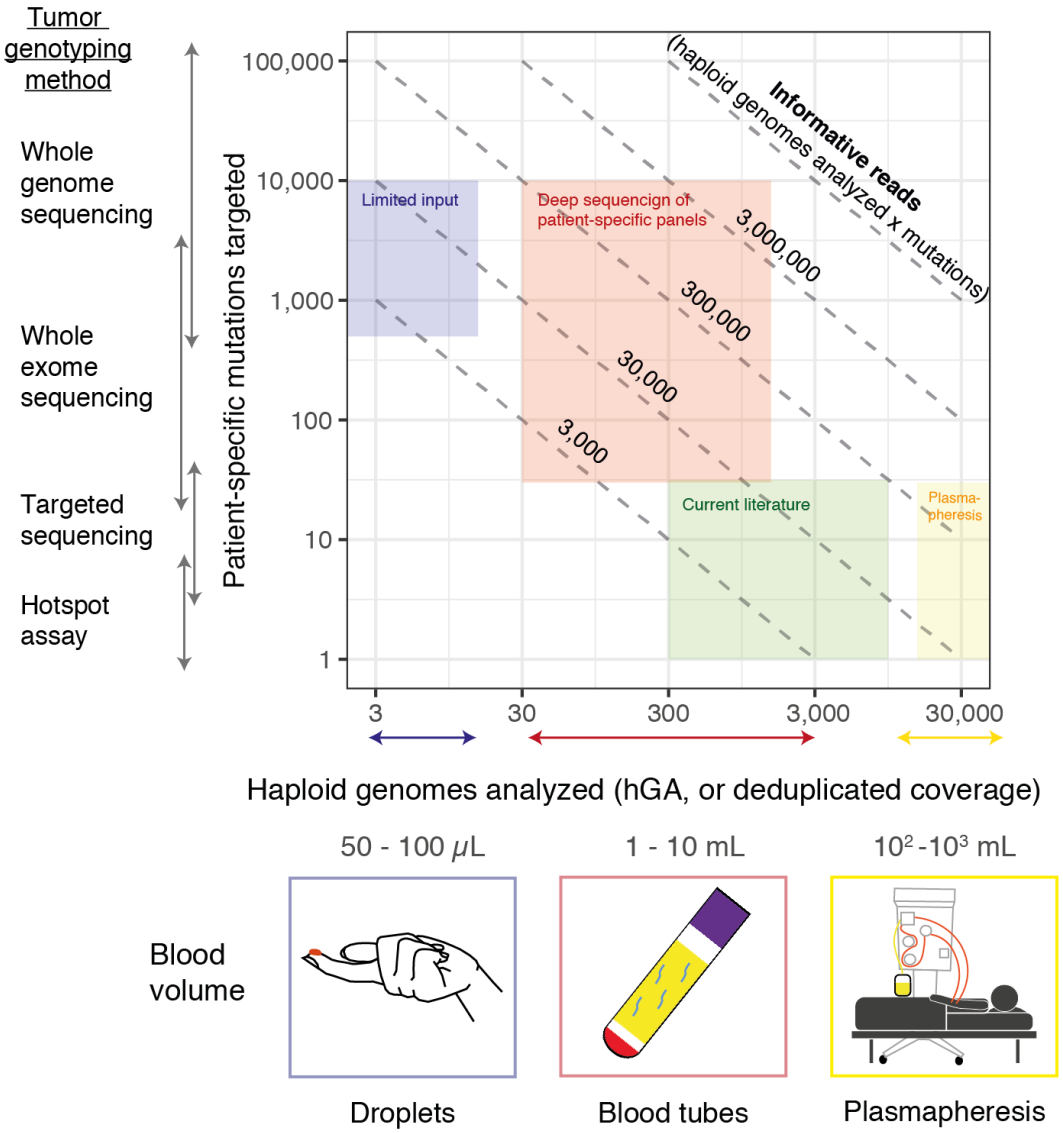


Fig. 5.1 **Two-dimensional representation of assay sensitivity for ctDNA detection.** ctDNA sensitivity can be viewed as a two-dimensional matrix where the product of DNA input and the number of targeted mutations determines the overall sensitivity of the assay.

sequencing of matched tumour tissue. Calling mutations from the entire genome would substantially increase the number of mutations identified in a given patient's tumour and also result in a greater sensitivity for ctDNA detection. However, given the current cost of sequencing, this approach is not feasible on a large scale. In future, tumour whole genome sequencing is expected to be performed more routinely, both in the research setting as well as in the clinic, and would provide the necessary sequencing data to perform more sensitive ctDNA detection.

5.2 Towards a simplified sample collection

Another major hurdle in the liquid biopsy field has been appropriate sample collection and processing. Samples are either collected in EDTA-containing tubes, requiring prompt spinning after collection or in more costly tubes with added preservatives/fixatives which allow for a more lenient processing [133]. If the sample is not processed correctly, any ctDNA signal will be diluted by genomic DNA, which is broken down and can hamper downstream analysis. Sample collection facilities have to ensure that they have trained staff and the required equipment to adhere to these standards. Samples are therefore usually only collected in large clinical centres with the required expertise and facilities to ensure appropriate processing. However, this may become restrictive to the frequency of sample collection. A simplified sample collection method could enable new trial designs and widen clinical applications.

In this PhD I explored the possibility of collecting samples in a way that could be universally applied to most clinical centres. By collecting blood spots one can avoid the immediate sample processing requirements described above. Furthermore, sample storage requirements are reduced considerably. One can envision a general practitioner collecting a blood spot from a patient, letting it dry for a few minutes and sending it by regular post to the nearest research facility, where the sample will be further processed. In the future, it could become possible for the patient to collect their own sample at home. Additionally, due to the minimal sampling volume and ease of collection, it becomes feasible to obtain multiple time points from the same patient, facilitating the exploration of additional research questions. My research on minimal sampling volume forms a proof of concept study, highlighting the detection of ctDNA from a blood spot in a sample from a patient with melanoma. I also applied the same method to a sample from an ovarian PDX model. Most traditional ctDNA analysis requires volumes that are only obtainable from small rodents through terminal bleeding of the animal, prohibiting longitudinal sampling. I was able to detect both cancerous ctDNA (coming from the engrafted tumour and aligning to the human genome) and non-

cancerous cfDNA (aligning to the mouse genome) from a PDX blood spot sample. As a tail vain blood spot can be obtained without culling the animal, this application could provide a novel means for the analysis of samples from small rodents. Further research and additional samples will have to be analysed to determine the robustness and sensitivity of this method.

In the future I hope to apply INVAR to blood spot data. If patient-specific tumour mutations are known, INVAR could be run straight from whole genome sequencing data as demonstrated in section 2.13. In that instance, ctDNA was detected at a sequencing depth of just 0.6x. In chapter 4 I obtain a sequencing depth of 6x for the blood spot, which would (given the same number of patient-specific mutations) result in a ten-fold greater sensitivity based on the sensitivity matrix in Fig. 5.1. As mentioned above, sensitivity can be increased even further by obtaining patient-specific mutations from whole genome sequencing of the matched tumour tissue: for example, maintaining a sequencing depth of 6x and analysing 10,000 patient specific mutations (realistic in whole genome sequencing of high mutation burden cancers), one would obtain a sensitivity of $1.7\text{x}10^{-5}$ when analysing a single drop of blood.

5.3 The future of circulating tumour DNA

One of the current areas of focus in the ctDNA field lies in the detection of minimal volumes of disease, either to diagnose cancer in early-stage such as undiagnosed or treatment naïve patients or to discover minimal residual disease in patients after completing treatment, aiming to identify relapse sooner.

By combining multiple cancer markers or utilising machine learning some assays have shown promising results in the detection of very low levels of disease. For example, the ctDNA based part of the CancerSeek assay showed only limited sensitivity in the cohort. Once eight additional protein markers were added, the detection rate of the assay increased between two- and ten-fold and resulted in detection of stage I disease in 43% of cases [98]. Similarly, two recent approaches utilise machine learning on sWGS data from plasma samples to detect cancer more sensitively. Machine learning enables the exploration of large datasets with manifold features and provides tools to reveal hidden structures and patterns in these data. Such patterns can then be leveraged to build a more robust method for discrimination of samples containing cancer from healthy control samples. Mouliere and colleagues show that using CNVs and fragmentation features in a machine learning approach improves the identification of cancerous samples compared to using the two parameters individually [27]. The second approach by Cristiano and colleagues train a model based on genome wide fragmentation features from sWGS data and distinguish cancer samples

from healthy controls [132]. All three approaches yield promising results in the detection of cancer without previous information on the tumour mutations and may pave the way for future detection assays that are being developed in this space.

While the field has improved in the detection of minimal residual disease, it is lacking large scale studies to identify cancer in the general population to generate a catalogue of potential markers for the early detection of cancer. GRAIL, a spin out company from Illumina, set up the Circulating Cell-free Genome Atlas Study (CCGA) in 2016 to characterise cancer signals in the blood from a population of 15,000 healthy and cancerous patients. In 2017 they launched the STRIVE study, which has recruited 100,000 women after taking a mammogram to study the detection of breast and other cancers using a blood-based methylation test that was developed as part of the CCGA effort. In 2019 the SUMMIT study was launched, a prospective observational study of 50,000 patients, some of which have a higher risk of developing lung and other cancers. Using the same methylation-based blood test, the population will be screened for the presence of cancer and the identification of the cancer type. Similarly, Guardant Health, a precision oncology company, has developed the LUNAR assay, which detects cancer using genomic alterations and epigenetic signatures. Currently, the assay has only been applied in the minimal residual disease setting but Guardant Health plans to apply the test for detection of early-stage cancer from the general population as well. Both GRAIL and Guardant Health are setting the path for the future of liquid biopsy and its potential use to detect cancer in the general population. The data they generate will be crucial to the development of a future population wide screening tool.

It should be noted that other than analysing components in the blood, direct sampling for cancer has also proven advantageous in its early detection and may provide yet another avenue to detect cancer population-wide in the future. The main drawback for direct sampling approaches is that they are only applicable to cancers with an accessible sample for analysis. Examples are urine for urological cancers, saliva for cancers of the oral cavity, cerebrospinal fluid for brain tumours, stool for colorectal cancers, cervical sampling for cervical cancer and cell sampling for oesophageal cancer. Some studies have compared direct sampling to ctDNA analysis from plasma and saw discordant detection in the cohort, indicating that information can be gained by analysing both the blood and direct samples where possible. In the future, accepting cost and practical considerations, any test should try and analyse a multitude of sample types and biomarkers within each sample to try and maximise overall detection.

While many of the methods discussed in this chapter are currently only used in the research setting, they nonetheless point towards the potential of future cancer diagnosis and monitoring. Together with the large scale efforts of GRAIL and Guardant Health, it

could become possible to identify seemingly healthy early-stage cancer patients amongst the total population, thereby increasing their chances of survival. Additionally, patients who underwent successful treatment will enter a longitudinal follow up where sensitive (patient-specific) methods (either based on liquid biopsies or direct sampling) are utilised to monitor the patient closely and minimally invasively to detect a potential relapse as early as possible, thereby increasing the chances of survival for the patient. Combined, these advances in cancer diagnosis and monitoring of minimally residual disease can substantially enhance the overall survival chances of the patients and become the future of cancer monitoring.

Chapter 6

Publications

6.1 Manuscripts

J. C. M. Wan*, **K. Heider***, D. Gale, S. Murphy, E. Fischer, J. Morris, F. Mouliere, D. Chandrananda, A. Marshall, A. B. Gill, P. Y. Chan, E. Barker, G. Young, W. N. Cooper, I. Hudecova, F. Marass, G. R. Bignell, C. Alifrangis, M. R. Middleton, F. A. Gallagher, C. Parkinson, A. Durrani, U. McDermott, C. G. Smith, C. Massie, P. G. Corrie, N. Rosenfeld, High-sensitivity monitoring of ctDNA by patient-specific sequencing panels and integration of variant reads, *bioRxiv* (2019); [Available online: https://doi.org/10.1101/759399]
* Joint first authors

K. Heider, F. Mouliere, C. G. Smith, Overview of selected approaches for cell free DNA library preparation and sequencing, as part of the book "Circulating Tumour DNA – Purification and Analysis Techniques", *World Scientific Publishing - to be published*.

K. Heider*, J. C. M. Wan*, J. Hall, S. Boyle, I. Hudecova, D. Gale, W. N. Cooper, P. G. Corrie, J. D. Brenton, C. G. Smith, N. Rosenfeld, Detection of ctDNA from dried blood spots after DNA size selection, *bioRxiv* (2019); [Available online: https://doi.org/10.1101/759365]
* Joint first authors

J. C. M. Wan*, **K. Heider***, D. Gale, S. Murphy, E. Fischer, F. Mouliere, A. Ruiz-Valdepenas, A. Santonja, J. Morris, D. Chandrananda, A. Marshall, A. B. Gill, P. Y. Chan, E. Barker, G. Young, W. N. Cooper, I. Hudecova, F. Marass, R. Mair, K. M. Brindle, G. D. Stewart, J. Abraham, C. Caldas, D. M. Rassl, R. C. Rintoul, G. R. Bignell, C. Alifrangis, M. R. Middleton, F. A. Gallagher, C. Parkinson, A. Durrani, U. McDermott, C. G. Smith, C. Massie, P. G. Corrie, N. Rosenfeld, ctDNA monitoring to parts per million using patient-specific sequencing and

integration of variant reads, *in preparation*

F. Mouliere, A. M. Piskorz, D. Chandrananda, E. Moore, J. Morris, C. G. Smith, T. Goranova, **K. Heider**, R. Mair, A. Supernat, I. Gounaris, S. Ros, J. C. M. Wan, M. Jimenez-Linan, D. Gale, K. Brindle, C. E. Massie, C. A. Parkinson, J. D. Brenton, N. Rosenfeld, Selecting Short DNA Fragments In Plasma Improves Detection Of Circulating Tumour DNA, *bioRxiv* (2017); [Available online: https://doi.org/10.1101/134437]

F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair, T. Goranova, F. Marass, **K. Heider**, J. C. M. Wan, A. Supernat, I. Hudecova, I. Gounaris, S. Ros, M. Jimenez-Linan, J. Garcia-Corbacho, K. Patel, O. Østrup, S. Murphy, M. D. Eldridge, D. Gale, G. D. Stewart, J. Burge, W. N. Cooper, M. S. van der Heijden, C. E. Massie, C. Watts, P. Corrie, S. Pacey, K. M. Brindle, R. D. Baird, M. Mau-Sørensen, C. A. Parkinson, C. G. Smith, J. D. Brenton, N. Rosenfeld, Enhanced detection of circulating tumor DNA by fragment size analysis, *Sci. Transl. Med.* **10**, eaat4921 (2018)

F. Mouliere, **K. Heider**, C. G. Smith, J. Su, M. Thompson, J. Morris, J. C. M. Wan, D. Chandrananda, J. Hadfield, M. Grezlak, I. Hudecova, W. Cooper, D. Gale, M. Eldridge, C. Watts, K. Brindle, N. Rosenfeld, R. Mair, Integrated clonal analysis reveals circulating tumor DNA in urine and plasma of glioma patients, *bioRxiv* (2019); [Available online: https://doi.org/10.1101/758441]

C. G. Smith, T. Moser, J. Burge, M. Eldridge, A. L. Riediger, F. Mouliere, D. Chandrananda, **K. Heider**, J. C. M. Wan, A. Y. Warren, J. Morris, I. Hudecova, W. N. Cooper, T. J. Mitchell, D. Gale, A. Ruiz-Valdepenas, T. Klatte, S. Ursprung, E. Sala, A. C. P. Riddick, T. F. Aho, J. N. Armitage, S. Perakis, M. Pichler, M. Seles, G. Wcislo, S. J. Welsh, A. Matakidou, T. Eisen, C. E. Massie, N. Rosenfeld, E. Heitzer, G. D. Stewart, Comprehensive characterisation of cell-free tumour DNA in plasma and urine of patients with renal tumours, *bioRxiv* (2019); [Available online: https://doi.org/10.1101/758003]

6.2 Abstracts

K. Heider, J. C. M. Wan, D. Gale, A. Ruiz-Valdepenas, F. Mouliere, J. Morris, W. Qian, J. Wulff, N. Demiris, A. Williams, T. Eisen, C. G. Smith, D. M. Rass, S. Harden, R. C. Rintoul, C. Massie, N. Rosenfeld, Improved ctDNA detection in early stage non-small cell lung cancer, *11th CNAPS Symposium in Jerusalem*

K. Heider, J. C. M. Wan, D. Gale, F. Mouliere, W. Qian, A. Kateb, G. Doughton, N. Ramenatte, R. Tysoe, C. G. Smith, D. M. Rassl, S. Harden, R. C. Rintoul, C. Massie, N. Rosenfeld, Improved ctDNA detection in early stage non-small-cell lung cancer, *AACR 2019 conference in Atlanta*

K. Heider, A. Ruiz-Valdepenas, G. Doughton, W. Qian, D. Chandrananda, C. G. Smith, E. Mosseley, T. Young, C. Castedo, A. Stone, T. Green, A. Walker, C. Thorbinson, M. Griffiths, D. Gale, C. E. Massie, T. Eisen, D. M. Rassl, S. Harden, R. C. Rintoul, N. Rosenfeld, ctDNA in early stage NSCLC (LUCID study): high sensitivity analysis in low burden disease, *10th CNAPS Symposium in Montpellier*

A. Ruiz-Valdepenas, **K. Heider**, G. Doughton, W. Qian, C. Massie, D. Chandrananda, C. Smith, D. Gale, E. Moseley, C. Castedo, A. Stone, C. Thorbinson, T. Eisen, D. Rassl, S. Harden, R. Rintoul, N. Rosenfeld, MA 11.02 Circulating Tumor DNA in Early Stage NSCLC: High Sensitivity Analysis in Low Burden Disease. LUCID Study Update, *J. Thorac. Oncol.* **12**, S1843–S1844 (2017).

6.3 Patents

Patent application 1819134.6: **Improvements in Variant Detection.**
Application filed in 2018 by Cancer Research Technologies Ltd.

6.4 Software packages

INVAR (not yet published) - https://bitbucket.org/nrlab/invar/wiki/Home

References

[1] P Mandel and P Metais. Les acides nucléiques du plasma sanguin chez l'homme. *C. R. Acad. Sci. Paris*, 142:241–243, feb 1948.

[2] Y. M. Dennis Lo, Noemi Corbetta, Paul F. Chamberlain, Vik Rai, Ian L. Sargent, Christopher W.G. Redman, and James S. Wainscoat. Presence of fetal DNA in maternal plasma and serum. *Lancet (London, England)*, 350(9076):485–7, aug 1997.

[3] Sarah Breitbach, Suzan Tug, and Perikles Simon. Circulating cell-free DNA: An up-coming molecular marker in exercise physiology. *Sports Medicine*, 42(7):565–586, 2012.

[4] Iwijn De Vlaminck, Hannah A Valantine, Thomas M Snyder, Calvin Strehl, Garrett Cohen, Helen Luikart, Norma F Neff, Jennifer Okamoto, Daniel Bernstein, Dana Weisshaar, Stephen R Quake, and Kiran K Khush. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine*, 6(241):241ra77, jun 2014.

[5] Edison Moraes Rodrigues Filho, Daniel Simon, Nilo Ikuta, Caroline Klovan, Fernando Augusto Dannebrock, Carla Oliveira de Oliveira, and Andrea Regner. Elevated cell-free plasma DNA level as an independent predictor of mortality in patients with severe traumatic brain injury. *Journal of neurotrauma*, 31(19):1639–46, oct 2014.

[6] S A Leon, B Shapiro, D M Sklaroff, and M J Yaros. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer research*, 37(3):646–50, mar 1977.

[7] Heidi Schwarzenbach, Dave S B Hoon, and Klaus Pantel. Cell-free nucleic acids as biomarkers in cancer patients. *Nature reviews. Cancer*, 11(6):426–437, 2011.

[8] Jonathan C. M. Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D. Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, 17(4):223–238, feb 2017.

[9] Y. M.Dennis Lo, K. C.Allen Chan, Hao Sun, Eric Z. Chen, Peiyong Jiang, Fiona M.F. Lun, Yama W. Zheng, Tak Y. Leung, Tze K. Lau, Charles R. Cantor, and Rossa W.K. Chiu. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science Translational Medicine*, 2(61), 2010.

[10] Dineika Chandrananda, Natalie P. Thorne, and Melanie Bahlo. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Medical Genomics*, 8(1):29, dec 2015.

[11] A V Lichtenstein, H S Melkonyan, L D Tomei, and S R Umansky. Circulating nucleic acids and apoptosis. *Annals of the New York Academy of Sciences*, 945:239–49, sep 2001.

[12] H. Christina Fan, Yair J. Blumenfeld, Usha Chitkara, Louanne Hudgins, and Stephen R. Quake. Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clinical Chemistry*, 56(8):1279–1286, 2010.

[13] Giulia Siravegna, Silvia Marsoni, Salvatore Siena, and Alberto Bardelli. Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology*, mar 2017.

[14] Florent Mouliere, Bruno Robert, Erika Arnau Peyrotte, Maguy Del Rio, Marc Ychou, Franck Molina, Celine Gongora, and Alain R. Thierry. High Fragmentation Characterizes Tumour-Derived Circulating DNA. *PLoS ONE*, 6(9):e23418, sep 2011.

[15] Hunter R. Underhill, Jacob O. Kitzman, Sabine Hellwig, Noah C. Welker, Riza Daza, Daniel N. Baker, Keith M. Gligorich, Robert C. Rostomily, Mary P. Bronner, and Jay Shendure. Fragment Length of Circulating Tumor DNA. *PLoS Genetics*, 12(7):e1006162, jul 2016.

[16] European Medicines Agency. Iressa: EPAR - Product Information. Technical report, 2014.

[17] K. C.Allen Chan, Peiyong Jiang, Yama W.L. Zheng, Gary J.W. Liao, Hao Sun, John Wong, Shing Shun N. Siu, Wing C. Chan, Stephen L. Chan, Anthony T.C. Chan, Paul B.S. Lai, Rossa W.K. Chiu, and Y. M.D. Lo. Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clinical Chemistry*, 59(1):211–224, jan 2013.

[18] Muhammed Murtaza, Sarah-Jane Dawson, Dana W Y Tsui, Davina Gale, Tim Forshew, Anna M Piskorz, Christine Parkinson, Suet-Feung Chin, Zoya Kingsbury, Alvin S C Wong, Francesco Marass, Sean Humphray, James Hadfield, David Bentley, Tan Min Chin, James D Brenton, Carlos Caldas, and Nitzan Rosenfeld. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108–12, 2013.

[19] Christopher Abbosh, Nicolai J. Birkbak, Gareth A. Wilson, Mariam Jamal-Hanjani, Tudor Constantin, Raheleh Salari, John Le Quesne, David A. Moore, Selvaraju Veeriah, Rachel Rosenthal, Teresa Marafioti, Eser Kirkizlar, Thomas B.K. Watkins, Nicholas McGranahan, Sophia Ward, Luke Martinson, Joan Riley, Francesco Fraioli, Maise Al Bakir, Eva Grönroos, Francisco Zambrana, Raymondo Endozo, Wenya Linda Bi, Fiona M. Fennessy, Nicole Sponer, Diana Johnson, Joanne Laycock, Seema Shafi, Justyna Czyzewska-Khan, Andrew Rowan, Tim Chambers, Nik Matthews, Samra Turajlic, Crispin Hiley, Siow Ming Lee, Martin D. Forster, Tanya Ahmad, Mary Falzon, Elaine Borg, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M. Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Dina Hafez, Ashwini Naik, Apratim Ganguly, Stephanie Kareht, Rajesh Shah, Leena Joseph, Anne Marie Quinn, Phil A. Crosbie, Babu Naidu, Gary

Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A. Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F. Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Dahmane Oukrif, Ayse U. Akarca, John A. Hartley, Helen L. Lowe, Sara Lock, Natasha Iles, Harriet Bell, Yenting Ngai, Greg Elgar, Zoltan Szallasi, Roland F. Schwarz, Javier Herrero, Aengus Stewart, Sergio A. Quezada, Karl S. Peggs, Peter Van Loo, Caroline Dive, C. Jimmy Lin, Matthew Rabinowitz, Hugo J.W.L. Aerts, Allan Hackshaw, Jacqui A. Shaw, Bernhard G. Zimmermann, the TRACERx consortium, the PEACE consortium, and Charles Swanton. Phylogenetic ctDNA analysis depicts early stage lung cancer evolution. *Nature*, 22364(7655):1–25, apr 2017.

[20] Reza Fazel, Harlan M Krumholz, Yongfei Wang, Joseph S Ross, Jersey Chen, Henry H Ting, Nilay D Shah, Khurram Nasir, Andrew J Einstein, and Brahmajee K Nallamothu. Exposure to low-dose ionizing radiation from medical imaging procedures. *The New England journal of medicine*, 361(9):849–57, aug 2009.

[21] Michael J. Overman, Janhavi Modak, Scott Kopetz, Ravi Murthy, James C. Yao, Marshall E. Hicks, James L. Abbruzzese, and Alda L. Tam. Use of research biopsies in clinical trials: Are risks and benefits adequately discussed? *Journal of Clinical Oncology*, 31(1):17–22, 2013.

[22] G Sozzi, K Musso, C Ratcliffe, P Goldstraw, M A Pierotti, and U Pastorino. Detection of microsatellite alterations in plasma DNA of non-small cell lung cancer patients: a prospect for early diagnosis. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 5(10):2689–92, oct 1999.

[23] Elizabeth M Swisher, Melissa Wollan, Sarita M Mahtani, Julia B Willner, Rochelle Garcia, Barbara A Goff, and Mary-Claire King. Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *American journal of obstetrics and gynecology*, 193(3 Pt 1):662–7, sep 2005.

[24] Maurice Stroun, Philippe Anker, Jacqueline Lyautey, Christine Lederrey, and Pierre A. Maurice. Isolation and characterization of DNA from the plasma of cancer patients. *European Journal of Cancer and Clinical Oncology*, 23(6):707–712, jun 1987.

[25] FDA. cobas® EGFR Mutation Test v2 approval.

[26] Elena Pereira, Olga Camacho-Vanegas, Sanya Anand, Robert Sebra, Sandra Catalina Camacho, Leopold Garnar-Wortzel, Navya Nair, Erin Moshier, Melissa Wooten, Andrew Uzilov, Rong Chen, Monica Prasad-Hayes, Konstantin Zakashansky, Ann Marie Beddoe, Eric Schadt, Peter Dottino, and John A. Martignetti. Personalized Circulating Tumor DNA Biomarkers Dynamically Predict Treatment Response and Survival In Gynecologic Cancers. *Plos One*, 10(12):e0145754, 2015.

[27] Florent Mouliere, Dineika Chandrananda, Anna M. Piskorz, Elizabeth K. Moore, James Morris, Lise Barlebo Ahlborn, Richard Mair, Teodora Goranova, Francesco Marass, Katrin Heider, Jonathan C. M. Wan, Anna Supernat, Irena Hudecova, Ioannis Gounaris, Susana Ros, Mercedes Jimenez-Linan, Javier Garcia-Corbacho, Keval Patel,

Olga Østrup, Suzanne Murphy, Matthew D. Eldridge, Davina Gale, Grant D. Stewart, Johanna Burge, Wendy N. Cooper, Michiel S. van der Heijden, Charles E. Massie, Colin Watts, Pippa Corrie, Simon Pacey, Kevin M. Brindle, Richard D. Baird, Morten Mau-Sørensen, Christine A. Parkinson, Christopher G. Smith, James D. Brenton, and Nitzan Rosenfeld. Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine*, 10(466):eaat4921, nov 2018.

[28] Christopher Abbosh, Nicolai J. Birkbak, and Charles Swanton. Early stage NSCLC — challenges to implementing ctDNA-based screening and MRD detection. *Nature Reviews Clinical Oncology*, 15(9):577–586, jul 2018.

[29] A Szpechcinski, J Chorostowska-Wynimko, R Struniawski, W Kupis, P Rudzinski, R Langfort, E Puscinska, P Bielen, P Sliwinski, and T Orlowski. Cell-free DNA levels in plasma of patients with non-small-cell lung cancer and inflammatory lung disease. *British journal of cancer*, 113(3):476–83, jul 2015.

[30] Chetan Bettegowda, Mark Sausen, Rebecca J Leary, Isaac Kinde, Yuxuan Wang, Nishant Agrawal, Bjarne R Bartlett, Hao Wang, Brandon Luber, Rhoda M Alani, Emmanuel S Antonarakis, Nilofer S Azad, Alberto Bardelli, Henry Brem, John L Cameron, Clarence C Lee, Leslie A Fecher, Gary L Gallia, Peter Gibbs, Dung Le, Robert L Giuntoli, Michael Goggins, Michael D Hogarty, Matthias Holdhoff, Seung-Mo Hong, Yuchen Jiao, Hartmut H Juhl, Jenny J Kim, Giulia Siravegna, Daniel A Laheru, Calogero Lauricella, Michael Lim, Evan J Lipson, Suely Kazue Nagahashi Marie, George J Netto, Kelly S Oliner, Alessandro Olivi, Louise Olsson, Gregory J Riggins, Andrea Sartore-Bianchi, Kerstin Schmidt, le Ming Shih, Sueli Mieko Oba-Shinjo, Salvatore Siena, Dan Theodorescu, Jeanne Tie, Timothy T Harkins, Silvio Veronese, Tian-Li Wang, Jon D Weingart, Christopher L Wolfgang, Laura D Wood, Dongmei Xing, Ralph H Hruban, Jian Wu, Peter J Allen, C Max Schmidt, Michael A Choti, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, Nickolas Papadopoulos, and Luis A Diaz. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science translational medicine*, 6(224):224ra24, 2014.

[31] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville C W Eclov, Leslie a Modlin, Chih Long Liu, Joel W Neal, Heather a Wakelee, Robert E Merritt, Joseph B Shrager, Billy W Loo, Ash a Alizadeh, and Maximilian Diehn. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*, 20(5):548–54, 2014.

[32] Christine A. Parkinson, Davina Gale, Anna M. Piskorz, Heather Biggs, Charlotte Hodgkin, Helen Addley, Sue Freeman, Penelope Moyle, Evis Sala, Karen Sayal, Karen Hosking, Ioannis Gounaris, Mercedes Jimenez-Linan, Helena M. Earl, Wendi Qian, Nitzan Rosenfeld, and James D. Brenton. Exploratory Analysis of TP53 Mutations in Circulating Tumour DNA as Biomarkers of Treatment Response for Patients with Relapsed High-Grade Serous Ovarian Carcinoma: A Retrospective Study. *PLOS Medicine*, 13(12):e1002198, dec 2016.

[33] Joshua Moss, Judith Magenheim, Daniel Neiman, Hai Zemmour, Netanel Loyfer, Amit Korach, Yaacov Samet, Myriam Maoz, Henrik Druid, Peter Arner, Keng-Yeh Fu, Endre Kiss, Kirsty L. Spalding, Giora Landesberg, Aviad Zick, Albert Grinshpun, A. M. James Shapiro, Markus Grompe, Avigail Dreazan Wittenberg, Benjamin Glaser,

Ruth Shemer, Tommy Kaplan, and Yuval Dor. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications*, 9(1):5068, dec 2018.

[34] Rebecca J Leary, Isaac Kinde, Frank Diehl, Kerstin Schmidt, Chris Clouser, Cisilya Duncan, Alena Antipova, Clarence Lee, Kevin McKernan, Francisco M De La Vega, Kenneth W Kinzler, Bert Vogelstein, Luis A Diaz, and Victor E Velculescu. Development of personalized tumor biomarkers using massively parallel sequencing. *Science translational medicine*, 2(20):20ra14, feb 2010.

[35] Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana W Y Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M Piskorz, Mercedes Jimenez-Linan, David Bentley, James Hadfield, Andrew P May, Carlos Caldas, James D Brenton, and Nitzan Rosenfeld. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med*, 4(136):136ra68, 2012.

[36] Aadel A Chaudhuri, Jacob J Chabon, Alexander F Lovejoy, Aaron M Newman, Henning Stehr, Tej D Azad, Michael S Khodadoust, Mohammad Shahrokh Esfahani, Chih Long Liu, Li Zhou, Florian Scherer, David M Kurtz, Carmen Say, Justin N Carter, David J Merriott, Jonathan C Dudley, Michael S Binkley, Leslie Modlin, Sukhmani K Padda, Michael F Gensheimer, Robert B West, Joseph B Shrager, Joel W Neal, Heather A Wakelee, Billy W Loo, Ash A Alizadeh, and Maximilian Diehn. Early Detection of Molecular Residual Disease in Localized Lung Cancer by Circulating Tumor DNA Profiling. *Cancer discovery*, 7(12):1394–1403, dec 2017.

[37] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, oct 2008.

[38] Sten Linnarsson. Recent advances in DNA sequencing methods – general principles of sample preparation. *Experimental Cell Research*, 316(8):1339–1343, 2010.

[39] Steven R Head, H Kiyomi Komori, Sarah A LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, 56(2):61–4, 66, 68, passim, 2014.

[40] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, sep 2014.

[41] Michael A Quail, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow, and Daniel J Turner. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12):1005–1010, dec 2008.

[42] Umberto Malapelle, Pasquale Pisapia, Danilo Rocco, Riccardo Smeraglio, Maria di Spirito, Claudio Bellevicine, and Giancarlo Troncone. Next generation sequencing techniques in liquid biopsy: focus on non-small cell lung cancer patients. *Translational lung cancer research*, 5(5):505–510, oct 2016.

[43] Van Dijk EL. et al. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, 322(1):12–20, 2014.

[44] Joaquim S L Vong, Jason C H Tsang, Peiyong Jiang, Wing-Shan Lee, Tak Yeung Leung, K C Allen Chan, Rossa W K Chiu, and Y M Dennis Lo. Single-Stranded DNA Library Preparation Preferentially Enriches Short Maternal DNA in Maternal Plasma. *Clinical chemistry*, page clinchem.2016.268656, mar 2017.

[45] Philip Burnham, Min Seong Kim, Sean Agbor-Enoh, Helen Luikart, Hannah A. Valantine, Kiran K. Khush, and Iwijn De Vlaminck. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Scientific Reports*, 6:27859, jun 2016.

[46] Marie-Theres Gansauge and Matthias Meyer. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4):737–748, mar 2013.

[47] Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 164(1):57–68, 2016.

[48] Marie-Theres Gansauge, Tobias Gerber, Isabelle Glocke, Petra Korlevic, Laurin Lippik, Sarah Nagel, Lara Maria Riehl, Anna Schmidt, and Matthias Meyer. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic acids research*, 45(10):e79, jun 2017.

[49] Florent Mouliere and Nitzan Rosenfeld. Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proceedings of the National Academy of Sciences of the United States of America*, 112(11):3178–9, mar 2015.

[50] Peiyong Jiang, Carol W M Chan, K C Allen Chan, Suk Hang Cheng, John Wong, Vincent Wai-Sun Wong, Grace L H Wong, Stephen L Chan, Tony S K Mok, Henry L Y Chan, Paul B S Lai, Rossa W K Chiu, and Y M Dennis Lo. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proceedings of the National Academy of Sciences of the United States of America*, 112(11):E1317–25, 2015.

[51] Tina Moser, Peter Ulz, Qing Zhou, Samantha Perakis, Jochen B. Geigl, Michael R. Speicher, and Ellen Heitzer. Single-stranded DNA library preparation does not preferentially enrich circulating tumor DNA. *Clinical Chemistry*, 63(10):1656–1659, 2017.

[52] Y Y Zhu, E M Machleder, A Chenchik, R Li, and P D Siebert. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques*, 30(4):892–7, apr 2001.

[53] Andrey Turchinovich, Harald Surowy, Andrius Serva, Marc Zapatka, Peter Lichter, and Barbara Burwinkel. Capture and Amplification by Tailing and Switching (CATS). An ultrasensitive ligation-independent method for generation of DNA libraries for deep sequencing from picogram amounts of DNA and RNA. *RNA biology*, 11(7):817–28, 2014.

[54] Oriya Vardi, Inbal Shamir, Elisheva Javasky, Alon Goren, and Itamar Simon. Biases in the SMART-DNA library preparation method associated with genomic poly dA/dT sequences. *PLOS ONE*, 12(2):e0172769, feb 2017.

[55] Alain R. Thierry, Florent Mouliere, Celine Gongora, Jeremy Ollier, Bruno Robert, Marc Ychou, Maguy Del Rio, and Franck Molina. Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic Acids Research*, 38(18):6159–6175, oct 2010.

[56] Florent Mouliere, Anna M. Piskorz, Dineika Chandrananda, Elizabeth Moore, James Morris, Christopher G. Smith, Teodora Goranova, Katrin Heider, Richard Mair, Anna Supernat, Ioannis Gounaris, Susana Ros, Jonathan C. M. Wan, Mercedes Jimenez-Linan, Davina Gale, Kevin Brindle, Charles E. Massie, Christine A. Parkinson, James D. Brenton, and Nitzan Rosenfeld. Selecting Short DNA Fragments In Plasma Improves Detection Of Circulating Tumour DNA. *bioRxiv*, 2017.

[57] D. I. Lou, J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences*, 110(49):19872–19877, dec 2013.

[58] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9530–5, jun 2011.

[59] Jérôme D. Robin, Andrew T. Ludlow, Ryan LaRanger, Woodring E. Wright, and Jerry W. Shay. Comparison of DNA Quantification Methods for Next Generation Sequencing. *Scientific Reports*, 6(1):24067, jul 2016.

[60] Hubert Hug and Rainer Schuler. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of theoretical biology*, 221(4):615–24, apr 2003.

[61] J. A. Casbon, R. J. Osborne, S. Brenner, and C. P. Lichtenstein. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, 39(12):e81–e81, jul 2011.

[62] Ruqin Kou, Ham Lam, Hairong Duan, Li Ye, Narisra Jongkam, Weizhi Chen, Shifang Zhang, and Shihong Li. Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLOS ONE*, 11(1):e0146638, jan 2016.

[63] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499, mar 2017.

[64] Jillian Phallen, Mark Sausen, Vilmos Adleff, Alessandro Leal, Carolyn Hruban, James White, Valsamo Anagnostou, Jacob Fiksel, Stephen Cristiano, Eniko Papp, Savannah Speir, Thomas Reinert, Mai-Britt Worm Orntoft, Brian D. Woodward, Derek Murphy, Sonya Parpart-Li, David Riley, Monica Nesselbush, Naomi Sengamalay, Andrew

Georgiadis, Qing Kay Li, Mogens Rørbæk Madsen, Frank Viborg Mortensen, Joost Huiskens, Cornelis Punt, Nicole van Grieken, Remond Fijneman, Gerrit Meijer, Hatim Husain, Robert B. Scharpf, Luis A. Diaz, Siân Jones, Sam Angiuoli, Torben Ørntoft, Hans Jørgen Nielsen, Claus Lindbjerg Andersen, and Victor E. Velculescu. Direct detection of early-stage cancers using circulating tumor DNA. *Science Translational Medicine*, 9(403), 2017.

[65] Aaron M. Newman, Alexander F. Lovejoy, Daniel M. Klass, David M. Kurtz, Jacob J. Chabon, Florian Scherer, Henning Stehr, Chih Long Liu, Scott V. Bratman, Carmen Say, Li Zhou, Justin N. Carter, Robert B. West, George W. Sledge, Joseph B. Shrager, Billy W. Loo, Joel W. Neal, Heather A. Wakelee, Maximilian Diehn, and Ash A. Alizadeh. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nature Biotechnology*, 34(5):547–555, 2016.

[66] Samantha Perakis and Michael R Speicher. Emerging concepts in liquid biopsies. *BMC medicine*, 15(1):75, apr 2017.

[67] University of Michigan. Connor - METHODS, 2016.

[68] Mikhail Shugay, Andrew R. Zaretsky, Dmitriy A. Shagin, Irina A. Shagina, Ivan A. Volchenkov, Andrew A. Shelenkov, Mikhail Y. Lebedin, Dmitriy V. Bagaev, Sergey Lukyanov, and Dmitriy M. Chudakov. MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLOS Computational Biology*, 13(5):e1005480, may 2017.

[69] Agilent. SureCall.

[70] Michael W Schmitt, Scott R Kennedy, Jesse J Salk, Edward J Fox, Joseph B Hiatt, and Lawrence A Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36):14508–13, sep 2012.

[71] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, may 2016.

[72] Tian-Li Wang, Christine Maierhofer, Michael R Speicher, Christoph Lengauer, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu. Digital karyotyping. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16156–61, dec 2002.

[73] Rebecca J. Leary, Mark Sausen, Isaac Kinde, Nickolas Papadopoulos, John D. Carpten, David Craig, Joyce O'Shaughnessy, Kenneth W. Kinzler, Giovanni Parmigiani, Bert Vogelstein, Luis A. Diaz, and Victor E. Velculescu. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Science Translational Medicine*, 4(162):162ra154, nov 2012.

[74] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, jan 2000.

[75] Viktor A. Adalsteinsson, Gavin Ha, Samuel S. Freeman, Atish D. Choudhury, Daniel G. Stover, Heather A. Parsons, Gregory Gydush, Sarah C. Reed, Denisse Rotem, Justin Rhoades, Denis Loginov, Dimitri Livitz, Daniel Rosebrock, Ignaty Leshchiner, Jaegil Kim, Chip Stewart, Mara Rosenberg, Joshua M. Francis, Cheng Zhong Zhang, Ofir Cohen, Coyin Oh, Huiming Ding, Paz Polak, Max Lloyd, Sairah Mahmud, Karla Helvie, Margaret S. Merrill, Rebecca A. Santiago, Edward P. O'Connor, Seong H. Jeong, Rachel Leeson, Rachel M. Barry, Joseph F. Kramkowski, Zhenwei Zhang, Laura Polacek, Jens G. Lohr, Molly Schleicher, Emily Lipscomb, Andrea Saltzman, Nelly M. Oliver, Lori Marini, Adrienne G. Waks, Lauren C. Harshman, Sara M. Tolaney, Eliezer M. Van Allen, Eric P. Winer, Nancy U. Lin, Mari Nakabayashi, Mary Ellen Taplin, Cory M. Johannessen, Levi A. Garraway, Todd R. Golub, Jesse S. Boehm, Nikhil Wagle, Gad Getz, J. Christopher Love, and Matthew Meyerson. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1):1324, dec 2017.

[76] Jelena Belic, Marina Koch, Peter Ulz, Martina Auer, Teresa Gerhalter, Sumitra Mohan, Katja Fischereder, Edgar Petru, Thomas Bauernhofer, Jochen B. Geigl, Michael R. Speicher, and Ellen Heitzer. Rapid identification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. *Clinical Chemistry*, 61(6):838–849, jun 2015.

[77] Ellen Heitzer, Peter Ulz, Jelena Belic, Stefan Gutschi, Franz Quehenberger, Katja Fischereder, Theresa Benezeder, Martina Auer, Carina Pischler, Sebastian Mannweiler, Martin Pichler, Florian Eisner, Martin Haeusler, Sabine Riethdorf, Klaus Pantel, Hellmut Samonigg, Gerald Hoefler, Herbert Augustin, Jochen B. Geigl, and Michael R. Speicher. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Medicine*, 5(4):30, 2013.

[78] Genohub. Whole Genome Sequencing (WGS) vs. Whole Exome Sequencing (WES), 2015.

[79] Eric Samorodnitsky, Benjamin M Jewell, Raffi Hagopian, Jharna Miya, Michele R Wing, Ezra Lyon, Senthilkumar Damodaran, Darshna Bhatt, Julie W Reeser, Jharna Datta, and Sameek Roychowdhury. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human mutation*, 36(9):903–14, sep 2015.

[80] Maria Romina Girotti, Gabriela Gremel, Rebecca Lee, Elena Galvani, Dominic Rothwell, Amaya Viros, Amit Kumar Mandal, Kok Haw Jonathan Lim, Grazia Saturno, Simon J. Furney, Franziska Baenke, Malin Pedersen, Jane Rogan, Jacqueline Swan, Matthew Smith, Alberto Fusi, Deemesh Oudit, Nathalie Dhomen, Ged Brady, Paul Lorigan, Caroline Dive, and Richard Marais. Application of Sequencing, Liquid Biopsies, and Patient-Derived Xenografts for Personalized Medicine in Melanoma. *Cancer Discovery*, 6(3), 2016.

[81] Steffen Dietz, Uwe Schirmer, Clémentine Mercé, Nikolas von Bubnoff, Edgar Dahl, Michael Meister, Thomas Muley, Michael Thomas, and Holger Sültmann. Low Input Whole-Exome Sequencing to Determine the Representation of the Tumor Exome in

Circulating DNA of Non-Small Cell Lung Cancer Patients. *PloS one*, 11(8):e0161012, 2016.

[82] Timothy M. Butler, Katherine Johnson-Camacho, Myron Peto, Nicholas J. Wang, Tara A. Macey, James E. Korkola, Theresa M. Koppie, Christopher L. Corless, Joe W. Gray, and Paul T. Spellman. Exome Sequencing of Cell-Free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. *PLOS ONE*, 10(8):e0136407, aug 2015.

[83] Sarah-Jane Dawson, Dana W.Y. Tsui, Muhammed Murtaza, Heather Biggs, Oscar M. Rueda, Suet-Feung Chin, Mark J. Dunning, Davina Gale, Tim Forshew, Betania Mahler-Araujo, Sabrina Rajan, Sean Humphray, Jennifer Becq, David Halsall, Matthew Wallis, David Bentley, Carlos Caldas, and Nitzan Rosenfeld. Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *New England Journal of Medicine*, 368(13):1199–1209, mar 2013.

[84] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, sep 2013.

[85] Cancer Research UK. Lung cancer statistics, 2016.

[86] Pengyuan Liu, Carl Morrison, Liang Wang, Donghai Xiong, Peter Vedell, Peng Cui, Xing Hua, Feng Ding, Yan Lu, Michael James, John D. Ebben, Haiming Xu, Alex A. Adjei, Karen Head, Jaime W. Andrae, Michael R. Tschannen, Howard Jacob, Jing Pan, Qi Zhang, Francoise Van den bergh, Haijie Xiao, Ken C. Lo, Jigar Patel, Todd Richmond, Mary Anne Watt, Thomas Albert, Rebecca Selzer, Marshall Anderson, Jiang Wang, Yian Wang, Sandra Starnes, Ping Yang, and Ming You. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, 33(7):1270–1276, 2012.

[87] Avrum Spira, Balazs Halmos, and Charles a Powell. Update in Lung Cancer 2014. *American journal of respiratory and critical care medicine*, 192(3):283–94, 2015.

[88] Eric A. Collisson, Joshua D. Campbell, Angela N. Brooks, Alice H. Berger, William Lee, Juliann Chmielecki, David G. Beer, Leslie Cope, Chad J. Creighton, Ludmila Danilova, Li Ding, Gad Getz, Peter S. Hammerman, D. Neil Hayes, Bryan Hernandez, James G. Herman, John V. Heymach, Igor Jurisica, Raju Kucherlapati, David Kwiatkowski, Marc Ladanyi, Gordon Robertson, Nikolaus Schultz, Ronglai Shen, Rileen Sinha, Carrie Sougnez, Ming Sound Tsao, William D. Travis, John N. Weinstein, Dennis A. Wigle, Matthew D. Wilkerson, Andy Chu, Andrew D. Cherniack, Angela Hadjipanayis, Mara Rosenberg, Daniel J. Weisenberger, Peter W. Laird, Amie Radenbaugh, Singer Ma, Joshua M. Stuart, Lauren Averett Byers, Stephen B. Baylin, Ramaswamy Govindan, Matthew Meyerson, Stacey B. Gabriel, Kristian Cibulskis, Jaegil Kim, Chip Stewart, Lee Lichtenstein, Eric S. Lander, Michael S. Lawrence, E. Getz, Robert Fulton, Lucinda L. Fulton, Michael D. McLellan, Richard K. Wilson, Kai Ye, Catrina C. Fronick, Christopher A. Maher, Christopher A. Miller, Michael C. Wendl, Christopher Cabanski, Elaine Mardis, David Wheeler, Miruna Balasundaram, Yaron S.N. Butterfield, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J.

Mungall, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Richard Varhol, A. Gordon Robertson, Natasja Wye, Nina Thiessen, Robert A. Holt, Steven J.M. Jones, Marco A. Marra, Marcin Imielinski, Robert C. Onofrio, Eran Hodis, Travis Zack, Elena Helman, Chandra Sekhar Pedamallu, Jill Mesirov, Gordon Saksena, Steven E. Schumacher, Scott L. Carter, Levi Garraway, Rameen Beroukhim, Semin Lee, Harshad S. Mahadeshwar, Angeliki Pantazi, Alexei Protopopov, Xiaojia Ren, Sahil Seth, Xingzhi Song, Jiabin Tang, Lixing Yang, Jianhua Zhang, Peng Chieh Chen, Michael Parfenov, Andrew Wei Xu, Netty Santoso, Lynda Chin, Peter J. Park, Katherine A. Hoadley, J. Todd Auman, Shaowu Meng, Yan Shi, Elizabeth Buda, Scot Waring, Umadevi Veluvolu, Donghui Tan, Piotr A. Mieczkowski, Corbin D. Jones, Janae V. Simons, Matthew G. Soloway, Tom Bodenheimer, Stuart R. Jefferys, Jeffrey Roach, Alan P. Hoyle, Junyuan Wu, Saianand Balu, Darshan Singh, Jan F. Prins, J. S. Marron, Joel S. Parker, Charles M. Perou, Jinze Liu, Dennis T. Maglinte, Philip H. Lai, Moiz S. Bootwalla, David J. Van Den Berg, Timothy Triche, Juok Cho, Daniel DiCara, David Heiman, Pei Lin, William Mallard, Douglas Voet, Hailei Zhang, Lihua Zou, Michael S. Noble, Nils Gehlenborg, Helga Thorvaldsdottir, Marc Danie Nazaire, Jim Robinson, B. Arman Aksoy, Giovanni Ciriello, Barry S. Taylor, Gideon Dresdner, Jianjiong Gao, Benjamin Gross, Venkatraman E. Seshan, Boris Reva, S. Onur Sumer, Nils Weinhold, Chris Sander, Sam Ng, Jingchun Zhu, Christopher C. Benz, Christina Yau, David Haussler, Paul T. Spellman, Patrick K. Kimes, Bradley M. Broom, Jing Wang, Yiling Lu, Patrick Kwok Shing Ng, Lixia Diao, Wenbin Liu, Christopher I. Amos, Rehan Akbani, Gordon B. Mills, Erin Curley, Joseph Paulauskis, Kevin Lau, Scott Morris, Troy Shelton, David Mallery, Johanna Gardner, Robert Penny, Charles Saller, Katherine Tarvin, William G. Richards, Robert Cerfolio, Ayesha Bryant, Daniel P. Raymond, Nathan A. Pennell, Carol Farver, Christine Czerwinski, Lori Huelsenbeck-Dill, Mary Iacocca, Nicholas Petrelli, Brenda Rabeno, Jennifer Brown, Thomas Bauer, Cureline Oleg Dolzhanskiy, Olga Potapova, Daniil Rotin, Olga Voronina, Elena Nemirovich-Danchenko, Konstantin V. Fedosenko, Anthony Gal, Madhusmita Behera, Suresh S. Ramalingam, Gabriel Sica, Douglas Flieder, Jeff Boyd, Jo Ellen Weaver, Bernard Kohl, Dang Huy Quoc Thinh, George Sandusky, Hartmut Juhl, Edwina Duhig, Peter Illei, Edward Gabrielson, James Shin, Beverly Lee, Kristen Rogers, Dante Trusty, Malcolm V. Brock, Christina Williamson, Eric Burks, Kimberly Rieger-Christ, Antonia Holway, Travis Sullivan, Michael K. Asiedu, Farhad Kosari, Natasha Rekhtman, Maureen Zakowski, Valerie W. Rusch, Paul Zippile, James Suh, Harvey Pass, Chandra Goparaju, Yvonne Owusu-Sarpong, John M.S. Bartlett, Sugy Kodeeswaran, Jeremy Parfitt, Harmanjatinder Sekhon, Monique Albert, John Eckman, Jerome B. Myers, Carl Morrison, Carmelo Gaudioso, Jeffrey A. Borgia, Philip Bonomi, Mark Pool, Michael J. Liptay, Fedor Moiseenko, Irina Zaytseva, Hendrik Dienemann, Michael Meister, Philipp A. Schnabel, Thomas R. Muley, Martin Peifer, Carmen Gomez-Fernandez, Lynn Herbert, Sophie Egea, Mei Huang, Leigh B. Thorne, Lori Boice, Ashley Hill Salazar, William K. Funkhouser, W. Kimryn Rathmell, Rajiv Dhir, Samuel A. Yousem, Sanja Dacic, Frank Schneider, Jill M. Siegfried, Richard Hajek, Mark A. Watson, Sandra McDonald, Bryan Meyers, Belinda Clarke, Ian A. Yang, Kwun M. Fong, Lindy Hunter, Morgan Windsor, Rayleen V. Bowman, Solange Peters, Igor Letovanec, Khurram Z. Khan, Mark A. Jensen, Eric E. Snyder, Deepak Srinivasan, Ari B. Kahn, Julien Baboud, David A. Pot, Kenna R.Mills Shaw, Margi Sheth, Tanja Davidsen, John A. Demchok, Liming Yang, Zhining Wang, Roy Tarnuzzer, Jean Claude Zenklusen, Bradley A. Ozenberger, Heidi J. Sofia, and Richard

Cheney. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, 511(7511):543–550, jul 2014.

[89] Charles Swanton and Ramaswamy Govindan. Clinical Implications of Genomic Discoveries in Lung Cancer. *New England Journal of Medicine*, 374(19):1864–1873, may 2016.

[90] Cancer Research UK. Incidence of common cancers (based on data provided by the Office of National Statistics, the Information Services Division Scotland, the Welsh Cancer Intelligence and Surveillance Unit, and the Northern Ireland Statistics and Research Agency), 2016.

[91] William D Travis, Elisabeth Brambilla, Andrew G Nicholson, Yasushi Yatabe, John H M Austin, Mary Beth Beasley, Lucian R Chirieac, Sanja Dacic, Edwina Duhig, Douglas B Flieder, Kim Geisinger, Fred R Hirsch, Yuichi Ishikawa, Keith M Kerr, Masayuki Noguchi, Giuseppe Pelosi, Charles A Powell, Ming Sound Tsao, Ignacio Wistuba, and WHO Panel. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 10(9):1243–1260, sep 2015.

[92] Close Window. Non – Small Cell Lung Cancer : Introduction, 2010.

[93] Medline plus. non small cell lung cancer, 2016.

[94] Cancer Research UK. Mortality of common cancers (based on data provided by the Office of National Statistics, the Information Services Division Scotland, and the Northern Ireland Statistics and Research Agency), 2016.

[95] Cancer Research UK. Cancer survival for common cancers, 2019.

[96] Marianne Bjerager, Torben Palshof, Ronald Dahl, Peter Vedsted, and Frede Olesen. Delay in diagnosis of lung cancer in general practice. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 56(532):863–8, nov 2006.

[97] W Hamilton, T J Peters, A Round, and D Sharp. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax*, 60(12):1059–1065, dec 2005.

[98] Joshua D. Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A. Javed, Fay Wong, Austin Mattox, Ralph. H. Hruban, Christopher L. Wolfgang, Michael G. Goggins, Marco Dal Molin, Tian-Li Wang, Richard Roden, Alison P. Klein, Janine Ptak, Lisa Dobbyn, Joy Schaefer, Natalie Silliman, Maria Popoli, Joshua T. Vogelstein, James D. Browne, Robert E. Schoen, Randall E. Brand, Jeanne Tie, Peter Gibbs, Hui-Li Wong, Aaron S. Mansfield, Jin Jen, Samir M. Hanash, Massimo Falconi, Peter J. Allen, Shibin Zhou, Chetan Bettegowda, Luis Diaz, Cristian Tomasetti, Kenneth W. Kinzler, Bert Vogelstein, Anne Marie Lennon, and Nickolas Papadopoulos. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, page eaar3247, jan 2018.

[99] Yali Xiong, Stacey Jeronis, Barbara Hoffman, Dan A. Liebermann, and Ossie Geifman-Holtzman. First trimester noninvasive fetal RHD genotyping using maternal dried blood spots. *Prenatal Diagnosis*, 37(4):311–317, 2017.

[100] Wei Luo, Hua Yang, Kimberly Rathbun, Chou Pong Pau, and Chin Yih Ou. Detection of human immunodeficiency virus type 1 DNA in dried blood spots by a duplex real-time PCR assay. *Journal of Clinical Microbiology*, 43(4):1851–1857, 2005.

[101] Rekha Gyanchandani, Erik Kvam, Ryan Heller, Erin Finehout, Nicholas Smith, Karthik Kota, John R. Nelson, Weston Griffin, Shannon Puhalla, Adam M. Brufsky, Nancy E. Davidson, and Adrian V. Lee. Whole genome amplification of cell-free DNA enables detection of circulating tumor DNA mutations from fingerstick capillary blood. *Scientific Reports*, 8(1):17313, dec 2018.

[102] Carlo Rago, David L. Huso, Frank Diehl, Baktiar Karim, Guosheng Liu, Nickolas Papadopoulos, Yardena Samuels, Victor E. Velculescu, Bert Vogelstein, Kenneth W. Kinzler, and Luis A. Diaz. Serial assessment of human tumor burdens in mice by the analysis of circulating DNA. *Cancer Research*, 67(19):9364–9370, oct 2007.

[103] Ignacio Varela, Patrick Tarpey, Keiran Raine, Dachuan Huang, Choon Kiat Ong, Helen Davies, David Jones, Meng-lay Lin, Jon Teague, Graham Bignell, Adam Butler, Juok Cho, Gillian L Dalgliesh, Danushka Galappaththige, Claire Hardy, Mingming Jia, Calli Latimer, King Wai Lau, John Marshall, Stuart Mclaren, Andrew Menzies, Laura Mudie, Lucy Stebbings, A David, L F a Wessels, Stephane Richard, Richard J Kahnoski, and John Anema. Exome sequencing identifies frequent mutation of the SWI / SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469(7331):539–542, 2011.

[104] Isaac Garcia-Murillas, Gaia Schiavon, Britta Weigelt, Charlotte Ng, Sarah Hrebien, Rosalind J. Cutts, Maggie Cheang, Peter Osin, Ashutosh Nerurkar, Iwanka Kozarewa, Javier Armisen Garrido, Mitch Dowsett, Jorge S. Reis-Filho, Ian E. Smith, and Nicholas C. Turner. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine*, 7(302), 2015.

[105] Jeanne Tie, Yuxuan Wang, Cristian Tomasetti, Lu Li, Simeon Springer, Isaac Kinde, Natalie Silliman, Mark Tacey, Hui-Li Wong, Michael Christie, Suzanne Kosmider, Iain Skinner, Rachel Wong, Malcolm Steel, Ben Tran, Jayesh Desai, Ian Jones, Andrew Haydon, Theresa Hayes, Tim J Price, Robert L Strausberg, Luis A Diaz Jr, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Peter Gibbs. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci. Transl. Med.*, 8(346):346ra92, 2016.

[106] R J Lee, G Gremel, A Marshall, K A Myers, N Fisher, J A Dunn, N Dhomen, P G Corrie, M R Middleton, P Lorigan, and R Marais. Circulating tumor DNA predicts survival in patients with resected high-risk stage II/III melanoma. *Annals of Oncology*, 29(2):490–496, feb 2018.

[107] Bradon R McDonald, Tania Contente-Cuomo, Stephen-John Sammut, Ahuva Odenheimer-Bergman, Brenda Ernst, Nieves Perdigones, Suet-Feung Chin, Maria Farooq, Patricia A Cronin, Karen S Anderson, Heidi E Kosiorek, Donald W Northfelt, Ann E McCullough, Bhavika K Patel, Carlos Caldas, Barbara A Pockaj, and

Muhammed Murtaza. Detection of residual disease after neoadjuvant therapy in breast cancer using personalized circulating tumor DNA analysis. *bioRxiv*, page 425470, sep 2018.

[108] Muhammed Murtaza, Sarah-Jane Dawson, Katherine Pogrebniak, Oscar M Rueda, Elena Provenzano, John Grant, Suet-Feung Chin, Dana W Y Tsui, Francesco Marass, Davina Gale, H Raza Ali, Pankti Shah, Tania Contente-Cuomo, Hossein Farahani, Karey Shumansky, Zoya Kingsbury, Sean Humphray, David Bentley, Sohrab P Shah, Matthew Wallis, Nitzan Rosenfeld, and Carlos Caldas. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nature communications*, 6:8760, 2015.

[109] Bradon R. McDonald, Tania Contente-Cuomo, Stephen-John Sammut, Ahuva Odenheimer-Bergman, Brenda Ernst, Nieves Perdigones, Suet-Feung Chin, Maria Farooq, Rosa Mejia, Patricia A. Cronin, Karen S. Anderson, Heidi E. Kosiorek, Donald W. Northfelt, Ann E. McCullough, Bhavika K. Patel, Jeffrey N. Weitzel, Thomas P. Slavin, Carlos Caldas, Barbara A. Pockaj, and Muhammed Murtaza. Personalized circulating tumor DNA analysis to detect residual disease after neoadjuvant therapy in breast cancer. *Science Translational Medicine*, 11(504):eaax7392, aug 2019.

[110] Clare Turnbull, Richard H. Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, Emma Baple, Clare Craig, Angela Hamblin, Shirley Henderson, Christine Patch, Amanda O'Neill, Andrew Devereaux, Katherine Smith, Antonio Rueda Martin, Alona Sosinsky, Ellen M. McDonagh, Razvan Sultana, Michael Mueller, Damian Smedley, Adam Toms, Lisa Dinh, Tom Fowler, Mark Bale, Tim Hubbard, Augusto Rendon, Sue Hill, and Mark J. Caulfield. The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Online)*, 361(April):1–7, 2018.

[111] K.C. Allen Chan, John K.S. Woo, Ann King, Benny C.Y. Zee, W.K. Jacky Lam, Stephen L. Chan, Sam W.I. Chu, Constance Mak, Irene O.L. Tse, Samantha Y.M. Leung, Gloria Chan, Edwin P. Hui, Brigette B.Y. Ma, Rossa W.K. Chiu, Sing-Fai Leung, Andrew C. van Hasselt, Anthony T.C. Chan, and Y.M. Dennis Lo. Analysis of Plasma Epstein–Barr Virus DNA to Screen for Nasopharyngeal Cancer. *New England Journal of Medicine*, 377(6):513–522, aug 2017.

[112] Maura Costello, Trevor J Pugh, Timothy J Fennell, Chip Stewart, Lee Lichtenstein, James C Meldrim, Jennifer L Fostel, Dennis C Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B Gabriel, Eric S Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research*, 41(6):e67, apr 2013.

[113] Rehan Akbani, Kadir C. Akdemir, B. Arman Aksoy, Monique Albert, Adrian Ally, Samirkumar B. Amin, Harindra Arachchi, Arshi Arora, J. Todd Auman, Brenda Ayala, Julien Baboud, Miruna Balasundaram, Saianand Balu, Nandita Barnabas, John Bartlett, Pam Bartlett, Boris C. Bastian, Stephen B. Baylin, Madhusmita Behera, Dmitry Belyaev, Christopher Benz, Brady Bernard, Rameen Beroukhim, Natalie Bir, Aaron D. Black, Tom Bodenheimer, Lori Boice, Genevieve M. Boland, Riccardo Bono, Moiz S. Bootwalla, Marcus Bosenberg, Jay Bowen, Reanne Bowlby, Christopher A.

Bristow, Laura Brockway-Lunardi, Denise Brooks, Jakub Brzezinski, Wiam Bshara, Elizabeth Buda, William R. Burns, Yaron S.N. Butterfield, Michael Button, Tiffany Calderone, Giancarlo Antonini Cappellini, Candace Carter, Scott L. Carter, Lynn Cherney, Andrew D. Cherniack, Aaron Chevalier, Lynda Chin, Juok Cho, Raymond J. Cho, Yoon La Choi, Andy Chu, Sudha Chudamani, Kristian Cibulskis, Giovanni Ciriello, Amanda Clarke, Stephen Coons, Leslie Cope, Daniel Crain, Erin Curley, Ludmila Danilova, Stefania D'Atri, Tanja Davidsen, Michael A. Davies, Keith A. Delman, John A. Demchok, Qixia A. Deng, Yonathan Lissanu Deribe, Noreen Dhalla, Rajiv Dhir, Daniel Dicara, Michael Dinikin, Michael Dubina, J. Stephen Ebrom, Sophie Egea, Greg Eley, Jay Engel, Jennifer M. Eschbacher, Konstantin V. Fedosenko, Ina Felau, Timothy Fennell, Martin L. Ferguson, Sheila Fisher, Keith T. Flaherty, Scott Frazer, Jessica Frick, Victoria Fulidou, Stacey B. Gabriel, Jianjiong Gao, Johanna Gardner, Levi A. Garraway, Julie M. Gastier-Foster, Carmelo Gaudioso, Nils Gehlenborg, Giannicola Genovese, Mark Gerken, Jeffrey E. Gershenwald, Gad Getz, Carmen Gomez-Fernandez, Thomas Gribbin, Jonna Grimsby, Benjamin Gross, Ranabir Guin, Tony Gutschner, Angela Hadjipanayis, Ruth Halaban, Benjamin Hanf, David Haussler, Lauren E. Haydu, D. Neil Hayes, Nicholas K. Hayward, David I. Heiman, Lynn Herbert, James G. Herman, Peter Hersey, Katherine A. Hoadley, Eran Hodis, Robert A. Holt, Dave Sb Hoon, Susan Hoppough, Alan P. Hoyle, Franklin W. Huang, Mei Huang, Sharon Huang, Carolyn M. Hutter, Matthew Ibbs, Lisa Iype, Anders Jacobsen, Valerie Jakrot, Alyssa Janning, William R. Jeck, Stuart R. Jefferys, Mark A. Jensen, Corbin D. Jones, Steven J.M. Jones, Zhenlin Ju, Hojabr Kakavand, Hyojin Kang, Richard F. Kefford, Fadlo R. Khuri, Jaegil Kim, John M. Kirkwood, Joachim Klode, Anil Korkut, Konstanty Korski, Michael Krauthammer, Raju Kucherlapati, Lawrence N. Kwong, Witold Kycler, Marc Ladanyi, Phillip H. Lai, Peter W. Laird, Eric Lander, Michael S. Lawrence, Alexander J. Lazar, Radoslaw Łaźniak, Darlene Lee, Jeffrey E. Lee, Junehawk Lee, Kenneth Lee, Semin Lee, William Lee, Ewa Leporowska, Kristen M. Leraas, Haiyan I. Li, Tara M. Lichtenberg, Lee Lichtenstein, Pei Lin, Shiyun Ling, Jia Liu, Ouida Liu, Wenbin Liu, Georgina V. Long, Yiling Lu, Singer Ma, Yussanne Ma, Andrzej Mackiewicz, Harshad S. Mahadeshwar, Jared Malke, David Mallery, Georgy M. Manikhas, Graham J. Mann, Marco A. Marra, Brenna Matejka, Michael Mayo, Sousan Mehrabi, Shaowu Meng, Matthew Meyerson, Piotr A. Mieczkowski, John P. Miller, Martin L. Miller, Gordon B. Mills, Fedor Moiseenko, Richard A. Moore, Scott Morris, Carl Morrison, Donald Morton, Stergios Moschos, Lisle E. Mose, Florian L. Muller, Andrew J. Mungall, Dawid Murawa, Pawel Murawa, Bradley A. Murray, Luigi Nezi, Sam Ng, Dana Nicholson, Michael S. Noble, Adeboye Osunkoya, Taofeek K. Owonikoko, Bradley A. Ozenberger, Elena Pagani, Oxana V. Paklina, Angeliki Pantazi, Michael Parfenov, Jeremy Parfitt, Peter J. Park, Woong Yang Park, Joel S. Parker, Francesca Passarelli, Robert Penny, Charles M. Perou, Todd D. Pihl, Olga Potapova, Victor G. Prieto, Alexei Protopopov, Michael J. Quinn, Amie Radenbaugh, Kunal Rai, Suresh S. Ramalingam, Ayush T. Raman, Nilsa C. Ramirez, Ricardo Ramirez, Uma Rao, W. Kimryn Rathmell, Xiaojia Ren, Sheila M. Reynolds, Jeffrey Roach, A. Gordon Robertson, Merrick I. Ross, Jason Roszik, Giandomenico Russo, Gordon Saksena, Charles Saller, Yardena Samuels, Chris Sander, Cindy Sander, George Sandusky, Netty Santoso, Melissa Saul, Robyn Pm Saw, Dirk Schadendorf, Jacqueline E. Schein, Nikolaus Schultz, Steven E. Schumacher, Charles Schwallier, Richard A. Scolyer, Jonathan Seidman, Pedamallu Chandra Sekhar, Harmanjatinder S. Sekhon, Yasin Senbabaoglu, Sahil

Seth, Kerwin F. Shannon, Samantha Sharpe, Norman E. Sharpless, Kenna R.Mills Shaw, Candace Shelton, Troy Shelton, Ronglai Shen, Margi Sheth, Yan Shi, Carolyn J. Shiau, Ilya Shmulevich, Gabriel L. Sica, Janae V. Simons, Rileen Sinha, Payal Sipahimalani, Heidi J. Sofia, Matthew G. Soloway, Xingzhi Song, Carrie Sougnez, Andrew J. Spillane, Arkadiusz Spychała, Jonathan R. Stretch, Joshua Stuart, Wiktoria M. Suchorska, Antje Sucker, S. Onur Sumer, Yichao Sun, Maria Synott, Barbara Tabak, Teresa R. Tabler, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Katherine Tarvin, Honorata Tatka, Barry S. Taylor, Marek Teresiak, Nina Thiessen, John F. Thompson, Leigh Thorne, Vesteinn Thorsson, Jeffrey M. Trent, Timothy J. Triche, Kenneth Y. Tsai, Peiling Tsou, David J. Van Den Berg, Eliezer M. Van Allen, Umadevi Veluvolu, Roeland G. Verhaak, Douglas Voet, Olga Voronina, Vonn Walter, Jessica S. Walton, Yunhu Wan, Yuling Wang, Zhining Wang, Scot Waring, Ian R. Watson, Nils Weinhold, John N. Weinstein, Daniel J. Weisenberger, Peter White, Matthew D. Wilkerson, James S. Wilmott, Lisa Wise, Maciej Wiznerowicz, Scott E. Woodman, Chang Jiun Wu, Chia Chin Wu, Junyuan Wu, Ye Wu, Ruibin Xi, Andrew Wei Xu, Da Yang, Liming Yang, Lixing Yang, Travis I. Zack, Jean C. Zenklusen, Hailei Zhang, Jianhua Zhang, Wei Zhang, Xiaobei Zhao, Jingchun Zhu, Kelsey Zhu, Lisa Zimmer, Erik Zmuda, and Lihua Zou. Genomic Classification of Cutaneous Melanoma. *Cell*, 2015.

[114] M. Jamal-Hanjani, G. A. Wilson, S. Horswell, R. Mitter, O. Sakarya, T. Constantin, R. Salari, E. Kirkizlar, S. Sigurjonsson, R. Pelham, S. Kareht, B. Zimmermann, and C. Swanton. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. *Annals of Oncology*, 27(5):862–867, may 2016.

[115] M. W. Schmitt, Scott R. Kennedy, Jesse J. Salk, Edward J. Fox, Joseph B. Hiatt, and Lawrence A. Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 109(36):14508–14513, 2007.

[116] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, jun 2013.

[117] P. G. Corrie, A. Marshall, P. D. Nathan, P. Lorigan, M. Gore, S. Tahir, G. Faust, C. G. Kelly, M. Marples, S. J. Danson, E. Marshall, S. J. Houston, R. E. Board, A. M. Waterston, J. P. Nobes, M. Harries, S. Kumar, A. Goodman, A. Dalgleish,

A. Martin-Clavijo, S. Westwell, R. Casasola, D. Chao, A. Maraveyas, P. M. Patel, C. H. Ottensmeier, D. Farrugia, A. Humphreys, B. Eccles, G. Young, E. O. Barker, C. Harman, M. Weiss, K. A. Myers, A. Chhabra, S. H. Rodwell, J. A. Dunn, and M. R. Middleton. Adjuvant bevacizumab for melanoma patients at high risk of recurrence: Survival analysis of the AVAST-M trial. *Annals of Oncology*, 29(8):1843–1852, 2018.

[118] Pippa G. Corrie, Andrea Marshall, Janet A. Dunn, Mark R. Middleton, Paul D. Nathan, Martin Gore, Neville Davidson, Steve Nicholson, Charles G. Kelly, Maria Marples, Sarah J. Danson, Ernest Marshall, Stephen J. Houston, Ruth E. Board, Ashita M. Waterston, Jenny P. Nobes, Mark Harries, Satish Kumar, Gemma Young, and Paul Lorigan. Adjuvant bevacizumab in patients with melanoma at high risk of recurrence (AVAST-M): Preplanned interim results from a multicentre, open-label, randomised controlled phase 3 study. *The Lancet Oncology*, 15(6):620–630, 2014.

[119] Rubicon. ThruPLEX ® Tag-seq Kit Instruction Manual Illumina® NGS Library Preparation with Unique Molecular Tags.

[120] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuze, Michael Soussan, Charlotte Robert, Claire Barakat, and Irene Buvat. A freeware for tumor heterogeneity characterization in PET, SPECT, CT, MRI and US to accelerate advances in radiomics. *Journal of Nuclear Medicine*, 58(supplement 1):1316, may 2017.

[121] Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620, 2014.

[122] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso Suárez, and Francisco Gude Sampedro. OptimalCutpoints : An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*, 61(8):1–36, 2014.

[123] Inivata, Cambridge, UK. Inivata web page, 2016.

[124] COSMIC. Comprehensive genomic characterization of squamous cell lung cancers, sep 2012.

[125] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo,

Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, aug 2013.

[126] Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S. Taylor, and Charles Swanton. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, dec 2016.

[127] Davina Gale, Andrew R. J. Lawson, Karen Howarth, Mikidache Madi, Bradley Durham, Sarah Smalley, John Calaway, Shannon Blais, Greg Jones, James Clark, Peter Dimitrov, Michelle Pugh, Samuel Woodhouse, Michael Epstein, Ana Fernandez-Gonzalez, Alexandra S. Whale, Jim F. Huggett, Carole A. Foy, Gerwyn M. Jones, Hadas Raveh-Amit, Karin Schmitt, Alison Devonshire, Emma Green, Tim Forshew, Vincent Plagnol, and Nitzan Rosenfeld. Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA. *PLOS ONE*, 13(3):e0194630, mar 2018.

[128] Ellen Heitzer, Imran S. Haque, Charles E. S. Roberts, and Michael R. Speicher. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, page 1, nov 2018.

[129] Jina Ko, Steven N Baldassano, Po-Ling Loh, Konrad Kording, Brian Litt, and David Issadore. Machine learning to detect signatures of disease in liquid biopsies - a user's guide. *Lab on a chip*, 18(3):395–405, 2018.

[130] Takara. Targeted Capture of ThruPLEX® Libraries with Agilent SureSelect®XT Target Enrichment System. Technical report.

[131] Matthew J. Wakefield. Xenomapper: Mapping reads in a mixed species context. *The Journal of Open Source Software*, 1(1):18, may 2016.

[132] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C. Bruhm, Sarah Østrup Jensen, Jamie E. Medina, Carolyn Hruban, James R. White, Doreen N. Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D. Woodward, Hatim Husain, Karlijn L. van Rooijen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J. H. van de Velde, Marcel Verheij, Annemieke Cats, Cornelis J. A. Punt, Geraldine R. Vink, Nicole C. T. van Grieken, Miriam Koopman, Remond J. A. Fijneman, Julia S. Johansen, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Robert B. Scharpf, and Victor E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, page 1, may 2019.

[133] Bente Risberg, Dana W Y Tsui, Heather Biggs, Andrea Ruiz-Valdepenas Martin de Almagro, Sarah-Jane Dawson, Charlotte Hodgkin, Linda Jones, Christine Parkinson, Anna Piskorz, Francesco Marass, Dineika Chandrananda, Elizabeth Moore, James Morris, Vincent Plagnol, Nitzan Rosenfeld, Carlos Caldas, James D Brenton, and Davina Gale. Effects of Collection and Processing Procedures on Plasma Circulating Cell-Free DNA from Cancer Patients. *The Journal of molecular diagnostics : JMD*, 20(6):883–892, nov 2018.

[134] Safia El Messaoudi, Fanny Rolet, Florent Mouliere, and Alain R. Thierry. Circulating cell free DNA: Preanalytical considerations. *Clinica Chimica Acta*, 424:222–230, sep 2013.

[135] Davina Gale, Vincent Plagnol, Andrew Lawson, Michelle Pugh, Sarah Smalley, Karen Howarth, Mikidache Madi, Bradley Durham, Vasudev Kumanduri, Kitty Lo, James Clark, Emma Green, Nitzan Rosenfeld, and Tim Forshew. Abstract 3639: Analytical performance and validation of an enhanced TAm-Seq circulating tumor DNA sequencing assay. In *Molecular and Cellular Biology, Genetics*, volume 76, pages 3639–3639. American Association for Cancer Research, jul 2016.

[136] Inc Beckman Coulter. SPRIselect User Guide.pdf.crdownload. *Beckman*, (October):1–30, 2012.

[137] Ji-Ping Wang. SPECIES: An R Package for Species Richness Estimation. *Journal of Statistical Software*, 40(9):1–15, 2011.

[138] Picard. Picard Metrics Definitions.